# HIGH-ORDER IMPLICIT MAXIMUM-PRINCIPLE-PRESERVING LOCAL DISCONTINUOUS GALERKIN METHODS FOR CONVECTION-DIFFUSION EQUATIONS

KAICHANG YU, JUAN CHENG, YUANYUAN LIU, AND CHI-WANG SHU

Abstract. We consider maximum-principle-preserving (MPP) property of two types of implicit local discontinuous Galerkin (LDG) schemes for solving diffusion and convection-diffusion equations. The first one is the original LDG scheme proposed in [B. Cockburn and C.-W. Shu, SIAM J. Numer. Anal. 35 (1998), pp.2440-2463] with backward Euler time discretization. The second one adds an MPP scaling limiter defined in [X. Zhang and C.-W. Shu, J. Comput. Phys., 229 (2010), pp.3091–3120] to the first one. Compared with explicit time discretization, implicit method allows for a larger time step. For pure diffusion equations in 1D, we prove that the second type of the LDG schemes is MPP, which can also achieve high order accuracy. This result can be generalized to 2D by using tensor product meshes but only for the second order $Q^1$ case. For convection-diffusion equations, the first type of LDG schemes, in the second order $P^1$ case in 1D, is proved to be MPP. In all the results above, in order to achieve the MPP property, it is necessary to have a lower bound on the time step in terms of the Courant-Friedrichs-Lewy (CFL) number. Although the analysis is only performed on linear equations, numerical experiments are provided to demonstrate that the second type of the LDG schemes works well in terms of the MPP property both for nonlinear convection-diffusion equations and for 2D higher order cases.

## 1. Introduction

In this paper, we consider the convection-diffusion equations with appropriate boundary conditions:

$$\begin{aligned} \partial_t u + \nabla \cdot \Phi(u) &= \nabla \cdot (\kappa(u)\nabla u), \ (\mathbf{x}, t) \in \Omega_T = \Omega \times (0, T], \\ u(\mathbf{x}, 0) &= u_0(\mathbf{x}), \qquad \mathbf{x} \in \Omega, \end{aligned} \tag{1.1}$$

where $\kappa(u)$ is a scalar function and satisfies $\kappa(u) \geq 0$ and $\Omega$ is a unit interval in 1D or a unit square in 2D, unless otherwise stated. For simplicity we mainly consider periodic boundary conditions. The exact solution of (1.1) always satisfies the maximum principle, i.e.,

$$u_0(\mathbf{x}) \in [m, M] \Longrightarrow u(\mathbf{x}, t) \in [m, M],$$

*Date*: May 23, 2024.

2020 *Mathematics Subject Classification.* Primary 65M60, 65M12.

*Key words and phrases.* local discontinuous Galerkin method, implicit time discretization, maximum-principle-preserving, convection-diffusion equations.

Research of the second author Juan Cheng is supported in part by National Key R&D Program of China No. 2023YFA1009003, and NSFC grant 12031001.

The third author Yuanyuan Liu is the corresponding author.

Research of the fourth author Chi-Wang Shu is supported in part by NSF grant DMS-2309249.

where $m = \min_{\mathbf{x}} u_0(\mathbf{x})$ and $M = \max_{\mathbf{x}} u_0(\mathbf{x})$. For numerical methods, we hope the numerical solution also satisfies the same maximum principle and call such methods maximum-principle-preserving (MPP), or positivity-preserving (PP) if the numerical solution only satisfies $u \geq 0$ specifically. These properties are important not only for numerical stability and robustness but also for making the numerical solution physically meaningful.

It is well-known that the discontinuous Galerkin (DG) method is a popular numerical method for convection and convection-diffusion equations as it has nice properties such as easy handling of complex geometry and high parallel efficiency, among others. The study of the DG methods includes those for the Runge-Kutta DG (RKDG) method for hyperbolic conservation laws [8, 9, 10, 11, 13, 14] and the local DG (LDG) method for diffusion and convection-diffusion equations [12, 15, 20, 34], among others.

In recent years, there has been a lot of attention paid to the MPP or PP properties of the DG method. They can be mainly divided into two categories. The first one is by firstly proving the cell averages of the unmodulated DG scheme satisfy the MPP or PP properties (this is often difficult and has to be analyzed for each types of equations and DG methods separately), then by applying a scaling limiter to achieve the MPP or PP properties for the DG solution without affecting the high order accuracy [35]. The methods in this category have the advantages of easy implementation, low computational cost, and mathematically guaranteed high order accuracy. Research for MPP/PP DG methods in this category includes those for hyperbolic conservation laws [7, 36, 37] and for the diffusion and convection-diffusion equations [16, 20, 38], among others. The other category of MPP/PP methods is based on parameterized flux limiters, such as those in [30]. The advantage of the DG methods in this category is that they are MPP/PP by design, however it is often difficult to rigorously prove that the resulting MPP/PP scheme is high order accurate. We refer interested readers to the review paper [32]. We remark that most of these works are based on explicit temporal discretizations, such as through strong stability preserving Runge-Kutta (SSPRK) methods or Lax-Wendroff type methods [33]. Explicit temporal discretizations have many advantages such as simplicity and low cost per time step without the need to solve large linear or nonlinear systems, easy to handle boundary conditions, high-order accuracy with SSP properties [18, 19] and so on. However, they suffer from time step restriction due to the Courant-Friedrichs-Lewy (CFL) constraints, especially for diffusion or convection-diffusion equations, which means that the time step $\tau$ must be restricted to have the same magnitude as the square of the spatial mesh size $h$. This restriction increases the global computational cost significantly, despite of the low cost per time step. In this paper, we consider implicit time discretization and mainly focus on the backward Euler method.

For continuous finite element methods, there have been many studies about the MPP property of implicit or time-independent scheme, such as [3, 23, 26, 28, 29, 31]. However, there are much fewer such studies for implicit DG schemes for solving diffusion or convection-diffusion equations. The first such result about the DG method is [21], where the authors derived a mesh condition to satisfy the MPP property of the $P^1$ interior penalty discontinuous Galerkin (IPDG) for 1D linear steady reaction-diffusion equations. Their basic approach is to prove that the matrix of the IPDG elliptic operator

is an $M$-matrix (see definitions in [4, 22]). This result can be applied directly to the $P^1$ IPDG scheme for solving 1D linear time-dependent diffusion equations with backward Euler time discretization and the mesh condition would be transformed to a lower bound of the time step by the CFL constraint. Later, Badia et al [1, 2] constructed and analyzed implicit IPDG schemes for two and three dimensions and for time-dependent problems. They made the schemes MPP through nonlinear stabilization and lumped mass. The schemes they proposed are at most second order accurate by numerical validation. As for higher order implicit DG schemes, van der Vegt et al [27] proposed an LDG scheme with diagonally implicit Runge-Kutta (DIRK) time discretization and the Karush-Kuhn-Tucker (KKT) limiter to get the MPP property. The usage of the KKT limiter requires the solution of a global optimization problem. Unlike for explicit schemes, the construction and analysis of the MPP/PP properties for implicit methods are more difficult, as the solution depends globally on the previous time step. It is necessary to study the specific structures of the matrices involved in the implicit solver in order to find sufficient conditions for the MPP/PP properties.

In this paper, we study the high order LDG scheme with backward Euler time discretization and consider its MPP property. In the following, when we say "implicit LDG scheme", it means that the LDG scheme with backward Euler time discretization specifically. The LDG method is first introduced by Cockburn and Shu in [12]. They rewrite (1.1) into a first-order system and use the DG scheme to solve it. Our first work is for any even order implicit LDG schemes for solving pure diffusion linear equations with periodic boundary condition in 1D, for which we use $x$ instead of $\mathbf{x}$. We prove that, if the solution at the current time level is MPP, then the cell averages obtained by the implicit LDG scheme at the next time level are MPP under a lower bound of the time step by the CFL condition. More specifically, it means that

$$u_h(x, t^n) \in [m, M] \Longrightarrow \bar{u}_h^{n+1} \in [m, M],$$

where $u_h(x, t^n)$ is the numerical solution at current time level $t^n$ and $\bar{u}_h^{n+1}$ represents the cell average of the numerical solution at the next time level $t^{n+1}$. Once this fact is established, we can add the scaling MPP limiter [35] to $u_h(x, t^{n+1})$ to get the MPP property. The technique of the proof is algebraic. For any even order schemes, we firstly find an implicit but clear expression of the cell averages by using the polynomial expansion of the $\delta$-function. Then we can obtain the MPP property by analyzing this special representation. For the third order scheme, it does not have the same representation, but thanks to the inverse property of special circulant matrices [6], we can obtain an explicit formula of the cell averages and subsequently we can get the conditions for the MPP property. Moreover, we extend the results to 2D but only for the second order $Q^1$ case. For all these analyses we assume uniform meshes. This construction and analysis for implicit MPP LDG schemes of the general even order and also the third order schemes are the main contributions of this paper and can be viewed as extensions of the work on implicit DG methods for solving hyperbolic conservation laws by Qin and Shu [24]. Compared with the work in [24], our problem is more complex because we are solving a first-order system rather than a scalar equation.

Our second work turns to consider convection-diffusion equations, which appears to be much more difficult than the individual convection or diffusion cases. For 1D linear

convection-diffusion equations with periodic boundary condition, we prove that the original implicit $P^1$ LDG scheme without any limiter is MPP under some mesh constraints and a lower bound on the time step. This proof is also algebraic, utilizes the properties of perturbed $M$-matrices, can handle certain nonuniform meshes and is an extension of Bouchon's work in [5].

To prepare for the next sections, we will introduce the implicit LDG scheme briefly.

### 1.1. **Implicit DG scheme for convection-diffusion equations.**

1.1.1. *The LDG discretization.* In this subsection, we present the LDG scheme for (1.1). The main idea of the LDG scheme is to rewrite (1.1) into the following first order system and apply the DG scheme to this system.

$$\begin{aligned} &\partial_t u + \sum_{1\le l\le \mathcal{D}} \partial_{x_l}\left(\Phi(u) - \sqrt{\kappa(u)}q_l\right) = 0, \quad \text{in } \Omega_T, \\ &q_l - \partial_{x_l} g_l(u) = 0, \qquad l = 1,\cdots,\mathcal{D}, \quad \text{in } \Omega_T, \end{aligned} \tag{1.2}$$

where $g_l(u) = \int^u \sqrt{\kappa(s)}ds$. The LDG method is then obtained by applying the DG method to (1.2). We refer the readers to [12] for more details and present below the 1D case as an example. First we decompose $\Omega = [0,1]$ into $N$ cells, $I_j = [x_{j-\frac{1}{2}}, x_{j+\frac{1}{2}}]$, for $j = 1, 2, \cdots, N$, and denote the size of the cell $I_j$ as $h_j = x_{j+\frac{1}{2}} - x_{j-\frac{1}{2}}$. The DG space is given as

$$V_h = \{v \in L^2(\Omega) : v|_{I_j} \in P^k(I_j),\ \forall j = 1,\cdots,N\},$$

where $P^k(I_j)$ denotes the space of polynomials with degree at most $k$ on $I_j$. Denote $(\cdot,\cdot)_j$ as the $L^2$ inner product on $I_j$. The semi-discrete LDG scheme for (1.2) is defined as follows, where (here and below) we abuse notation and denote the numerical solution of the DG scheme also as $u$ instead of the usual notation $u_h$. The scheme is: find $u, q \in V_h$, such that, for all $v, w \in V_h$ and any $j = 1,\cdots,N$, we have

$$\frac{d}{dt}(u(t), v)_j = \mathcal{B}_j^1(u, q; v), \quad \mathcal{B}_j^2(u, q; w) = 0, \tag{1.3}$$

where

(1.4)

$$\begin{aligned} \mathcal{B}_j^1(u,q;v) =& (\Phi(u) - \sqrt{\kappa(u)}q, v_x)_j + \left(\widehat{\sqrt{\kappa(u)}}\hat{q} - \widehat{\Phi(u)}\right)_{j+\frac{1}{2}} v^-_{j+\frac{1}{2}} - \left(\widehat{\sqrt{\kappa(u)}}\hat{q} - \widehat{\Phi(u)}\right)_{j-\frac{1}{2}} v^+_{j-\frac{1}{2}}, \\ \mathcal{B}_j^2(u,q;w) =& (q,w)_j + (g(u), w_x)_j - \left(\widehat{g(u)}_{j+\frac{1}{2}} w^-_{j+\frac{1}{2}} - \widehat{g(u)}_{j-\frac{1}{2}} w^+_{j-\frac{1}{2}}\right). \end{aligned}$$

The values $v^+_{j+\frac{1}{2}}$ and $v^-_{j+\frac{1}{2}}$ are the right and the left limits of the function $v(x)$ at the cell boundary $x_{j+\frac{1}{2}}$. The hat terms $\widehat{\Phi(u)}, \widehat{\sqrt{\kappa(u)}}, \widehat{g(u)}, \hat{q}$ are the numerical fluxes, where the diffusion fluxes are chosen as the alternating fluxes:

$$\widehat{\sqrt{\kappa(u)}} = \frac{g(u^+) - g(u^-)}{u^+ - u^-}, \quad \hat{q} = q^-, \quad \widehat{g(u)} = g(u^+), \tag{1.5}$$

and the convection flux is the Lax-Friedrichs flux, i.e.,

$$\widehat{\Phi(u)}_{j+\frac{1}{2}} = \frac{1}{2}\left(\Phi(u^-_{j+\frac{1}{2}}) + \Phi(u^+_{j+\frac{1}{2}}) - \mu\left(u^+_{j+\frac{1}{2}} - u^-_{j+\frac{1}{2}}\right)\right), \quad \mu = \max_u \left|\Phi'(u)\right|.$$

1.1.2. *Time discretization.* The backward Euler method is used to discretize the semi-discrete LDG scheme (1.3) in time. The numerical solution at time $t^{n+1}$, denoted by $u^{n+1}$, is obtained by finding $u^{n+1}, q \in V_h$, such that for all $v, w \in V_h$, we have

$$\frac{1}{\tau}(u^{n+1} - u^n, v)_j = \mathcal{B}_j^1(u^{n+1}, q; v), \qquad \mathcal{B}_j^2(u^{n+1}, q; w) = 0, \tag{1.6}$$

where $\tau = t^{n+1} - t^n$ is the time step.

1.1.3. *Definition of the maximum-principle-preserving (MPP) property.* Here we give the definition of the maximum-principle-preserving (MPP) property of numerical schemes.

**Definition 1.1.** *For given $u^n(x) \in [m, M],\ \forall x \in \Omega$, if we have $u^{n+1}(x) \in [m, M],\ \forall x \in \Omega$, where we recall that $m = \min_x u_0(x)$ and $M = \max_x u_0(x)$, then we call the scheme to be maximum-principle-preserving (MPP).*

We will make minor modifications to this definition later to make it easier to implement, by replacing $\forall x \in \Omega$ to $\forall x \in S$, where $S$ consists of suitable Legendre Gauss-Lobatto (LGL) quadrature points in each cell of $\Omega$.

1.2. **Organization of this paper.** The organization of this paper is as follows. In Section 2, we present the construction of high order implicit maximum-principle-preserving LDG scheme for the pure diffusion linear equation. To illustrate our results more clearly, for the one dimensional equation, we will use the cases of $k = 3$ and $k = 2$ as examples to provide a detailed explanation, and then will introduce a general theorem for any even order schemes. Furthermore, we extend this result to the $Q^1$ case in 2D. In Section 3, we give a proof of the second order implicit LDG scheme for 1D linear convection-diffusion equations. Numerical experiments for both linear and nonlinear equations are shown in Section 4 and further discussion is included in Section 5.

## 2. High order maximum-principle-preserving implicit LDG scheme for pure diffusion linear equations

In this section, we consider the construction and analysis of implicit high order maximum-principle-preserving LDG scheme for solving pure diffusion linear equations which means that

$$\Phi = 0,\ \kappa = 1,\ \Omega = [0, 1] \text{ or } [0, 1]^2, \tag{2.1}$$

in (1.1) with periodic boundary condition. Let us first briefly review the framework to achieve the MPP property introduced in [35]. Here we take 1D case as an example. Denote $\bar{u}_j$ as the cell average of $u(x)$ in the interval $I_j = [x_{j-1/2}, x_{j+1/2}]$, where $j = 1, \cdots, N$, and $\hat{I}$ as the reference cell $[-1, 1]$ and $T_j(x) = 2(x - x_j)/(x_{j+\frac{1}{2}} - x_{j-\frac{1}{2}})$ as the affine mapping between $I_j$ and $\hat{I}$, where $x_j = (x_{j-\frac{1}{2}} + x_{j+\frac{1}{2}})/2$. Also we denote $S_j$ as the set of the $k + 1$ LGL quadrature points in $I_j$. To get an MPP high-order DG scheme, the following two steps are followed in [35].

**Step 1.** Given $u^n(x) \in [m, M],\ \forall x \in S_j$, for any $j = 1, \cdots, N$. Then we need to prove that the cell average $\bar{u}_j^{n+1} \in [m, M]$ for all $j$ under certain time step restriction.

**Step 2.** Use a simple scaling limiter to modify the DG polynomial $u^{n+1}(x)$ in the cell $I_j$ into $\tilde{u}^{n+1}(x)$, such that $\tilde{u}^{n+1}(x) \in [m, M],\ \forall x \in S_j$, for any $j$, without changing the cell average and without affecting the high order accuracy. Then we take $\tilde{u}^{n+1}(x)$ as our numerical solution at the time level $n + 1$.

The first step is difficult to prove for our situation because $\bar{u}_j^{n+1}$ depends on $u^n(x)$ globally when we use implicit temporal discretization. Once we have $\bar{u}_j^{n+1} \in [m, M],\ \forall j$, then the second step follows from [35]. In the following we focus on the proof of $\bar{u}_j^{n+1} \in [m, M]$.

We follow the technique of proof introduced by Qin and Shu in [24]. First, we introduce the so-called $\delta_y^k-$polynomial with the definition

$$\delta_y^k(x) = \frac{1}{2}\sum_{l=0}^{k}(2l+1)p_l(x)p_l(y),\ x, y \in \hat{I}, \tag{2.2}$$

where $p_k(x)$ is the $k$-th Legendre polynomial. The polynomial $\delta_y^k(x)$ has similar properties to the Dirac delta distribution in $P^k(\hat{I})$, and we list more details in Appendix A.

In the following we will introduce some notations. We omit the superscript $k$ in $\delta_y^k$ and denote

$$\delta_{y,j}(x) = \frac{2}{h_j}\delta_y(T_j(x)),\ \text{where } y \in \hat{I} \text{ and } x \in I_j,\ j = 1, \cdots, N,$$

and $h_j = x_{j+\frac{1}{2}} - x_{j-\frac{1}{2}}$. We will use a uniform mesh unless otherwise stated, i.e., $h_j = h$ for all $j = 1, \cdots, N$. We define $\sigma = \tau/h^2$ and the identity matrix as $\mathbf{I}$, and define the circulant matrix $\mathbf{G} = C(0, 1, 0, \cdots, 0)$ as below:

$$\mathbf{G} = \begin{pmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ 0 & 0 & 0 & \ddots & 0 \\ 0 & 0 & 0 & \cdots & 1 \\ 1 & 0 & 0 & \cdots & 0 \end{pmatrix}. \tag{2.3}$$

2.1. **The MPP property of the implicit LDG scheme in 1D.** In this section, we assume that the degree $k$ of polynomials in the DG space is a fixed positive integer. The physical space $\Omega = [0, 1]$ in (1.1). Here and below, we shall denote the numerical solution $u^n$ at the current time step as $f$, and the numerical solution $u^{n+1}$ at the next time step as $u$. Then the LDG scheme in (1.4) becomes: find $u, q \in V_h$, such that

$$\mathcal{B}_j^1(u, q; v) = -(q, v_x)_j + \left(q_{j+\frac{1}{2}}^- v_{j+\frac{1}{2}}^- - q_{j-\frac{1}{2}}^- v_{j-\frac{1}{2}}^+\right) = \frac{(u, v)_j - (f, v)_j}{\tau}, \tag{2.4}$$

$$\mathcal{B}_j^2(u, q; w) = (q, w)_j + (u, w_x)_j - \left(u_{j+\frac{1}{2}}^+ w_{j+\frac{1}{2}}^- - u_{j-\frac{1}{2}}^+ w_{j-\frac{1}{2}}^+\right) = 0, \tag{2.5}$$

holds for all $v, w \in V_h$, where we recall that $\tau$ is the time step. Next we would like to prove that the cell average $\bar{u}_j$ lies in $[m, M]$, once $f(x)$ lies in $[m, M]$ for any $x \in S_j$ and for any $j = 1, \cdots, N$. For simplicity and clarity, we will mainly demonstrate the details when $k = 3$, and also show similar results for the specific cases of $k = 1, 5, 7, 9$. Then a general theorem for any odd $k$ is provided. The proof of the general theorem will be deferred to Appendix B. Notice that the specific theorems for $k = 1, 3, 5, 7, 9$ yield sharper lower bounds of $\sigma$ compared to the general theorem. The comparison is presented in Table 2.2.

2.1.1. *The MPP property of the LDG schemes with 2nd, 4th, 6th, 8th and 10th order accuracy.* In this subsection, we prove the MPP theorem for the fourth-order implicit LDG scheme, which can be extended to higher even orders. Firstly, we define the concept of the "periodic tridiagonal" matrices.

**Definition 2.1.** *If a circulant matrix has only diagonals, two adjacent non-diagonal lines, and the two elements at the top-right and bottom-left corners of this matrix are non-zero, then we call the matrix "periodic tridiagonal".*

It is clear that when we say a matrix $\mathbf{A} \in \mathcal{R}^{N\times N}$ is "periodic tridiagonal", it is equivalent that the matrix $\mathbf{A}$ can be written in the following form:

$$\mathbf{A} = a_1\mathbf{I} + a_2\mathbf{G} + a_3\mathbf{G}^{N-1}.$$

Then we shall prove the following theorem.

**Theorem 2.1.** *If we take the degree of polynomials $k = 3$, then we can rewrite the implicit LDG scheme in the following matrix form:*

$$\mathbf{M}\bar{\mathbf{u}} = \mathbf{Q}_1\mathbf{f}^+ + \mathbf{Q}_2\mathbf{f}_{-\sqrt{5}} + \mathbf{Q}_3\mathbf{f}_{\sqrt{5}} + \mathbf{Q}_4\mathbf{f}^-, \tag{2.6}$$

*where the matrices $\mathbf{M}$ and $\mathbf{Q}_i$ $(i = 1, 2, 3, 4)$ are "periodic tridiagonal". Recall that $\sigma = \tau/h^2$, then the definitions of these matrices are as follows:*

$$\mathbf{Q}_1 = 5\sigma\mathbf{D}^{-1}(4\mathbf{B}^\top\mathbf{T}_4 - \mathbf{A}\mathbf{T}_1) + 40\sigma\mathbf{T}_4 + \frac{1}{12}\mathbf{T}_1 + \mathbf{T}_4, \tag{2.7a}$$

$$\mathbf{Q}_2 = \frac{5-5\sqrt{5}}{2}\sigma\mathbf{D}^{-1}(4\mathbf{B}^\top\mathbf{T}_4 - \mathbf{A}\mathbf{T}_1) + (35\sqrt{5}+25)\sigma\mathbf{T}_4 + \frac{5}{12}\mathbf{T}_1, \tag{2.7b}$$

$$\mathbf{Q}_3 = \frac{5+5\sqrt{5}}{2}\sigma\mathbf{D}^{-1}(4\mathbf{B}^\top\mathbf{T}_4 - \mathbf{A}\mathbf{T}_1) + (-35\sqrt{5}+25)\sigma\mathbf{T}_4 + \frac{5}{12}\mathbf{T}_1, \tag{2.7c}$$

$$\mathbf{Q}_4 = 10\sigma\mathbf{D}^{-1}(4\mathbf{B}^\top\mathbf{T}_4 - \mathbf{A}\mathbf{T}_1) - 30\sigma\mathbf{T}_4 + \frac{1}{12}\mathbf{T}_1, \tag{2.7d}$$

*and*

$$\mathbf{M} = \mathbf{D}^{-1}(\mathbf{T}_1\mathbf{T}_3 - \mathbf{T}_2\mathbf{T}_4), \tag{2.8}$$

*where*

$$\mathbf{A} = \mathbf{I} - \mathbf{G}^{N-1},\ \mathbf{B} = (1+90\sigma)\mathbf{I} + (4+120\sigma)\mathbf{G},\ \mathbf{D} = (1+120\sigma)\mathbf{I} - 60\sigma\mathbf{G}^{N-1}.$$

*Also*

$$\begin{aligned}
&\mathbf{T}_1 = 840\sigma^2\mathbf{D}\mathbf{A}^\top + \mathbf{D}\mathbf{D}^\top + 16\sigma\mathbf{B}\mathbf{B}^\top = t_{11}\mathbf{I} + t_{12}\mathbf{G} + t_{13}\mathbf{G}^{N-1},\\
&\mathbf{T}_2 = 3360\sigma^2\mathbf{B}^\top = t_{21}\mathbf{I} + t_{22}\mathbf{G} + t_{23}\mathbf{G}^{N-1},\\
&\mathbf{T}_3 = \mathbf{D} + 840\sigma^2\mathbf{A} = t_{31}\mathbf{I} + t_{32}\mathbf{G} + t_{33}\mathbf{G}^{N-1},\\
&\mathbf{T}_4 = 4\sigma\mathbf{A}\mathbf{B} = t_{41}\mathbf{I} + t_{42}\mathbf{G} + t_{43}\mathbf{G}^{N-1}.
\end{aligned} \tag{2.10}$$

*where*

$t_{11} = 1 + 512\sigma + 37080\sigma^2 + 511200\sigma^3$, $\quad t_{12} = 4\sigma - 360\sigma^2 + 72000\sigma^3$,

$t_{13} = 4\sigma + 480\sigma^2 + 122400\sigma^3$, $\quad t_{21} = 3360\sigma^2 + 302400\sigma^3$, $\quad t_{22} = 0$,

$t_{23} = 13440\sigma^2 + 403200\sigma^3$, $\quad t_{31} = 1 + 120\sigma + 840\sigma^2$, $\quad t_{32} = 0$, $\quad t_{33} = -(60\sigma + 840\sigma^2)$,

$t_{41} = -12\sigma - 120\sigma^2$, $\quad t_{42} = 16\sigma + 480\sigma^2$, $\quad t_{43} = -(4\sigma + 360\sigma^2)$.

*The vectors are defined as below:*

$\bar{\mathbf{u}} = (\bar{u}_1, \cdots, \bar{u}_N)^\top$, $\mathbf{f}^+ = (f^+_{\frac{1}{2}}, \cdots, f^+_{N-\frac{1}{2}})^\top$, $\mathbf{f}^- = (f^-_{\frac{3}{2}}, \cdots, f^-_{N+\frac{1}{2}})^\top$, $\mathbf{f}_\alpha = (f_{1,\alpha}, \cdots, f_{N,\alpha})^\top$,

*where* $\alpha = \pm\sqrt{5}$, $f_{j,\alpha} := f(x_j \pm \frac{h}{2\alpha})$ *and* $\bar{u}_j = \frac{1}{|I_j|}\int_{I_j} u(x)dx$.

**Proof:** To get the cell averages, we take $v = 1$ in (1.6), then we have

$$\bar{u}_j = \bar{f}_j + \frac{\tau}{h}\left(q^-_{j+\frac{1}{2}} - q^-_{j-\frac{1}{2}}\right). \tag{2.12}$$

To eliminate the unknown terms $q^-$, we take $w = \delta_{1,j}(x)$ in (1.6), then we have

$$q^-_{j+\frac{1}{2}} + (u, \delta'_{1,j})_j - \left(u^+_{j+\frac{1}{2}}\delta_{1,j}(x_{j+\frac{1}{2}}) - u^+_{j-\frac{1}{2}}\delta_{1,j}(x_{j-\frac{1}{2}})\right) = 0, \tag{2.13}$$

where $\delta'_{1,j}(x)$ is the first derivative of the function $\delta_{1,j}(x)$. To eliminate $\delta'_{1,j}(x)$, we need to take $v(x) = \delta'_{1,j}(x)$ and $w(x) = \delta''_{1,j}(x)$ in (1.6) again. Repeat this operation until the third derivative of $\delta_{1,j}(x)$ is obtained. Here we omit some routine steps. Notice that

$$\begin{aligned}
&\delta_{\pm1,j}(x_{j\mp\frac{1}{2}}) = -\frac{4}{h}, \quad \delta'_{\pm1,j}(x_{j\mp\frac{1}{2}}) = \pm\frac{60}{h^2}, \quad \delta''_{\pm1,j}(x_{j\mp\frac{1}{2}}) = -\frac{360}{h^3}, \ \delta'''_{\pm1,j} = \pm\frac{840}{h^4},\\
&\delta_{\pm1,j}(x_{j\pm\frac{1}{2}}) = \frac{16}{h}, \quad \delta'_{\pm1,j}(x_{j\pm\frac{1}{2}}) = \pm\frac{120}{h^2}, \quad \delta''_{\pm1,j}(x_{j\pm\frac{1}{2}}) = \frac{480}{h^3},
\end{aligned}$$

where $\delta'_{\pm1,j}(x_{j\mp\frac{1}{2}}) = \pm\frac{60}{h^2}$ means $\delta'_{1,j}(x_{j-\frac{1}{2}}) = \frac{60}{h^2}$ and $\delta'_{-1,j}(x_{j+\frac{1}{2}}) = -\frac{60}{h^2}$ etc.. Finally we get

$$q^-_{j+\frac{1}{2}} = \left(\frac{4}{h} + \frac{360\tau}{h^3}\right)u^+_{j-\frac{1}{2}} + \left(\frac{16}{h} + \frac{480\tau}{h^3}\right)u^+_{j+\frac{1}{2}} - \frac{840\tau}{h^3}\bar{u}_j + \frac{60\tau}{h^2}\left(q^-_{j-\frac{1}{2}} - 2q^-_{j+\frac{1}{2}}\right) - (f, \delta'_{1,j})_j. \tag{2.15}$$

Similarly we have

$$\begin{aligned}
u^+_{j-\frac{1}{2}} = &-\left(\frac{4}{h} + \frac{360\tau}{h^3}\right)q^-_{j+\frac{1}{2}} - \left(\frac{16}{h} + \frac{480\tau}{h^3}\right)q^-_{j-\frac{1}{2}} + \frac{840\tau}{h^3}\bar{q}_j\\
&+ \frac{60\tau}{h^2}\left(u^+_{j+\frac{1}{2}} - 2u^+_{j-\frac{1}{2}}\right) + \tau(f, \delta''_{-1,j})_j + f^+_{j-\frac{1}{2}}.
\end{aligned} \tag{2.16}$$

Furthermore, we get the following equations in the matrix form:

$$\bar{\mathbf{u}} = \frac{\tau}{h}\mathbf{A}\mathbf{q}^- + \bar{\mathbf{f}}, \quad \bar{\mathbf{q}} = -\frac{1}{h}\mathbf{A}^\top \mathbf{u}^+, \quad \mathbf{D}\mathbf{q}^- = \frac{4}{h}\mathbf{B}\mathbf{u}^+ - \frac{840\sigma}{h}\bar{\mathbf{u}} - \mathbf{f}_1, \tag{2.17a}$$

$$\mathbf{D}^\top \mathbf{u}^+ = -\frac{4\tau}{h}\mathbf{B}^\top \mathbf{q}^- + 840\sigma\frac{\tau}{h}\bar{\mathbf{q}} + \mathbf{f}^+ + \tau\mathbf{f}_2, \tag{2.17b}$$

where

$$\mathbf{f}_1 = ((f, \delta'_{1,1})_1, \cdots, (f, \delta'_{1,N})_N)^\top, \ \mathbf{f}_2 = ((f, \delta''_{-1,1})_1, \cdots, (f, \delta''_{-1,N})_N)^\top, \tag{2.18}$$

and the definitions of $\mathbf{q}^-, \mathbf{u}^+, \bar{\mathbf{q}}$ or $\bar{\mathbf{f}}$ are similar to that of $\mathbf{f}^-, \mathbf{f}^+, \bar{\mathbf{u}}$ respectively. Based on the above equations, we can obtain the following equation through elimination:

$$(\mathbf{T}_1\mathbf{T}_3 - \mathbf{T}_2\mathbf{T}_4)\bar{\mathbf{u}} = (4\frac{\tau}{h}\mathbf{B}^\top\mathbf{T}_4 - \frac{\tau}{h}\mathbf{A}\mathbf{T}_1)\mathbf{f}_1 + \mathbf{D}\mathbf{T}_1\bar{\mathbf{f}} + \mathbf{D}\mathbf{T}_4\mathbf{f}^+ + \tau\mathbf{D}\mathbf{T}_4\mathbf{f}_2. \tag{2.19}$$

Notice that $\bar{\mathbf{f}}, \mathbf{f}_1, \mathbf{f}_2$ are all integrals of polynomials whose degrees are at most five, hence we can use the numerical quadrature with four LGL points to represent them exactly. Rewrite the cell average in (2.19) by point values at the LGL points, and multiply both sides of the equation by $\mathbf{D}^{-1}$, we obtain the equation (2.1). By the von Neumann expansion, we have

$$\mathbf{D}^{-1} = \frac{1}{1+120\sigma}\left(\mathbf{I} - \frac{60\sigma}{1+120\sigma}\mathbf{G}\right)^{-1} = \frac{1}{(1+120\sigma)(1-d^N)}\sum_{s=0}^{N-1} d^s\mathbf{G}^{N-s}, \tag{2.20}$$

where $d = \frac{60\sigma}{1+120\sigma}$. Denote $\mathbf{P} = 4\mathbf{B}^\top\mathbf{T}_4 - \mathbf{A}\mathbf{T}_1$. Notice that $\mathbf{T}_1$ and $\mathbf{T}_4$ are "periodic tridiagonal". Hence, to obtain our theorem, we only need to prove $\mathbf{M}$ and $\mathbf{D}^{-1}\mathbf{P}$ are "periodic tridiagonal". It is easy to see that:

(2.21)

$$\begin{aligned}\mathbf{P} =& (-1 - 300\sigma - 26880\sigma^2 - 252000\sigma^3)\mathbf{I} + (60\sigma + 8040\sigma^2 + 100800\sigma^3)\mathbf{G} + \\ & (1 + 300\sigma + 26040\sigma^2 + 201600\sigma^3)\mathbf{G}^{N-1} + (-60\sigma - 7200\sigma^2 - 50400\sigma^3)\mathbf{G}^{N-2} \\ \triangleq& \tilde{p}_1\mathbf{I} + \tilde{p}_2\mathbf{G} + \tilde{p}_3\mathbf{G}^{N-1} + \tilde{p}_4\mathbf{G}^{N-2}.\end{aligned}$$

Then we have

$$\begin{aligned}\mathbf{D}^{-1}\mathbf{P} &= \frac{1}{(1+120\sigma)(1-d^N)}\sum_{s=0}^{N-1} d^s\mathbf{G}^{N-s}\left(\tilde{p}_1\mathbf{I} + \tilde{p}_2\mathbf{G} + \tilde{p}_3\mathbf{G}^{N-1} + \tilde{p}_4\mathbf{G}^{N-2}\right) \\ &= \frac{1}{(1+120\sigma)(1-d^N)}\left(\tilde{p}_1\sum_{s=0}^{N-1} d^s + \tilde{p}_2\sum_{s=-1}^{N-2} d^{s+1} + \tilde{p}_3\sum_{s=1}^{N} d^{s-1} + \tilde{p}_4\sum_{s=2}^{N+1} d^{s-2}\right)\mathbf{G}^{N-s}.\end{aligned}$$

Notice that, for any fixed $s = 2, \cdots, N-2$, the coefficient of $\mathbf{G}^{N-s}$ is

$$\frac{d^{s-2}}{(1+120\sigma)(1-d^N)}(\tilde{p}_1 d^2 + \tilde{p}_2 d^3 + \tilde{p}_3 d + \tilde{p}_4) = 0.$$

Therefore we have

$$\mathbf{D}^{-1}\mathbf{P} = \tilde{\tilde{p}}_1\mathbf{I} + \tilde{\tilde{p}}_2\mathbf{G} + \tilde{\tilde{p}}_3\mathbf{G}^{N-1}, \tag{2.22}$$

where

$$\tilde{\tilde{p}}_1 = \frac{1}{(1+120\sigma)(1-d^N)}(\tilde{p}_1 + \tilde{p}_2 d + \tilde{p}_3 d^{N-1} + \tilde{p}_4 d^{N-2}) = -1 - 180\sigma - 1680\sigma^2,$$

and similarly

$$\tilde{\tilde{p}}_2 = 60\sigma(1+14\sigma),\ \tilde{\tilde{p}}_3 = 1 + 120\sigma + 840\sigma^2.$$

Notice that the above equation is independent of $N$. Also, it is clear that the matrices $\mathbf{T}_1$ and $\mathbf{T}_4$ are "periodic tridiagonal". Hence we have shown the matrices $\mathbf{Q}_i$ $(i = 1, 2, 3, 4)$ are all "periodic tridiagonal". Similarly, we have

$$\mathbf{M} = m_1\mathbf{I} + m_2\mathbf{G} + m_3\mathbf{G}^{N-1}, \tag{2.23}$$

where

$$m_1 = 1411200\sigma^4 + 662400\sigma^3 + 37920\sigma^2 + 512\sigma + 1, \tag{2.24}$$

$$m_2 = m_3 = -4\sigma\left(-1 + 90\sigma - 5400\sigma^2 + 176400\sigma^3\right). \tag{2.25}$$

Therefore, we finish the proof of this theorem. □

The above theorem is based on the compactness of the implicit LDG scheme itself. Fortunately, for $k = 3$, we can further obtain the three properties described in the following theorem, and thus obtain the MPP property.

**Theorem 2.2.** *Suppose* $\mathbf{f} \in [m, M]$, *where* $\mathbf{f} = \mathbf{f}^{\pm}$ *or* $\mathbf{f}_{\pm\sqrt{5}}$, *which means that each element of* $\mathbf{f}$ *lies in* $[m, M]$. *The matrix form* (2.6) *of the implicit LDG scheme enjoys the following three properties:*

*(i)* $\mathbf{M}$ *is an* $M$*-matrix when* $\sigma > 0.0194$.

*(ii) If* $\sigma > 0.0544$, *then* $\mathbf{Q}_i$ $(i = 1, 2, 3, 4)$, *are all positive matrices which means that their elements are all non-negative and at least one of them is positive.*

*(iii) Define* $\mathbf{W} = \sum_{i=1}^{4} \mathbf{Q}_i$. *Then we conclude that all row sums of the matrix* $\mathbf{M}$ *are equal and equal to all row sums of* $\mathbf{W}$.

*Notice that the cell average vector* $\bar{\mathbf{u}}$ *is a linear combination of* $\mathbf{f}^{\pm}$ *and* $\mathbf{f}_{\pm\sqrt{5}}$. *The first two properties above maintain the positivity of the combination coefficients. The third property ensures that the linear combination is convex. Furthermore, we conclude* $\bar{\mathbf{u}} \in [m, M]$ *when* $\sigma > 0.0544$.

**Proof:** Firstly, we prove that $\mathbf{M}$ is an $M$-matrix. By the definition, we only need to prove that $m_1 > 0$, $m_2 < 0$ and $m_1 > 2|m_2|$. According to (2.23) and (2.24), it is clear that $m_1 > 0$ and $m_1 > 2|m_2|$ as long as $\sigma > 0$. Furthermore, we obtain $m_2 < 0$ when $\sigma > 0.0194$ by numerically calculating the zeros of special polynomials.

Next, we prove the second property. Followed by (2.7) (2.10) and (2.22), we have

$$\mathbf{Q}_i = \tilde{q}_{i1}\mathbf{I} + \tilde{q}_{i2}\mathbf{G} + \tilde{q}_{i3}\mathbf{G}^{N-1},\ i = 1, 2, 3, 4.$$

For $\mathbf{Q}_1$, we have

$$\tilde{q}_{11} = \frac{1}{12} + \frac{77}{3}\sigma + 1590\sigma^2 + 29400\sigma^3,$$
$$\tilde{q}_{12} = \frac{1}{3}\sigma(49 + 4170\sigma + 88200\sigma^2),\ \tilde{q}_{13} = \frac{4}{3}\sigma(1 + 90\sigma).$$

It is clear that $\mathbf{Q}_1 > 0$ (i.e., each element of the matrix is non-negative and at least one of them is positive or $\mathbf{Q}_1$ is positive for simplicity) when $\sigma > 0$. For $\mathbf{Q}_2$, we have

$$\begin{aligned}
\tilde{q}_{21} &= \frac{5}{12}\left(1 + (506 + 6\sqrt{5})\sigma + 72(490 + \sqrt{5})\sigma^2 + 493920\sigma^3\right), \\
\tilde{q}_{22} &= \frac{5}{3}\sigma\left(1 + 6(40 + 41\sqrt{5})\sigma + 8820(3 + \sqrt{5})\sigma^2\right), \\
\tilde{q}_{23} &= -\frac{5}{6}\sigma\left(-5 + 3\sqrt{5} + 48(-10 + 11\sqrt{5})\sigma + 17640(\sqrt{5} - 3)\sigma^2\right).
\end{aligned}$$

It is easy to see that $\tilde{q}_{21},\ \tilde{q}_{22}$ are positive when $\sigma > 0$ and $\tilde{q}_{23}$ is positive if

$$\sigma > 0.0544. \tag{2.26}$$

For $\mathbf{Q}_3$, we have

$$\begin{aligned}
\tilde{q}_{31} &= \frac{5}{12}\left(1 + (506 - 6\sqrt{5})\sigma + 72(490 - \sqrt{5})\sigma^2 + 493920\sigma^3\right), \\
\tilde{q}_{32} &= -\frac{5}{3}\sigma\left(-1 + 6(-40 + 41\sqrt{5})\sigma + 8820(-3 + \sqrt{5})\sigma^2\right), \\
\tilde{q}_{33} &= \frac{5}{6}\sigma\left(5 + 3\sqrt{5} + 48(10 + 11\sqrt{5})\sigma + 17640(\sqrt{5} + 3)\sigma^2\right).
\end{aligned}$$

It is clear that $\tilde{q}_{31},\ \tilde{q}_{33}$ are positive when $\sigma > 0$ and $\tilde{q}_{32}$ is positive when

$$\sigma > 0.0426. \tag{2.27}$$

For $\mathbf{Q}_4$, we have

$$\begin{aligned}
\tilde{q}_{41} =& \frac{1}{12} + \frac{98}{3}\sigma + 1650\sigma^2 + 29400\sigma^3,\ \tilde{q}_{42} = \frac{1}{3}\sigma(1 + 270\sigma), \\
\tilde{q}_{43} =& \frac{1}{3}\sigma(31 + 4080\sigma + 88200\sigma^2).
\end{aligned}$$

It is easy to see that these elements of $\mathbf{Q}_4$ are positive when $\sigma > 0$. By combining (2.26) and (2.27), we get the second property.

It is easy to check all row sums of matrix $\mathbf{M}$ are equal and equal to all row sums of matrix $\mathbf{W}$. The summation is

$$705600\sigma^3 + 37200\sigma^2 + 520\sigma + 1.$$

Now we have the three properties. Next we prove the MPP property of the cell average. It is clear that the first two properties ensure that the coefficients of this combination are all positive. Then we shall prove that the summation of these coefficients is exactly one. Divide both sides of the equation by $705600\sigma^3 + 37200\sigma^2 + 520\sigma + 1$ and denote the corresponding newly obtained matrix as $\widetilde{\mathbf{M}} = (\tilde{m}_{ij})_{N\times N}$ and $\widetilde{\mathbf{W}} = (\tilde{w}_{ij})_{N\times N}$, where the row sums of the matrices $\widetilde{\mathbf{M}}$ and $\widetilde{\mathbf{W}}$ are exactly one. Then the sum of each row of the matrix $\widetilde{\mathbf{M}}^{-1} = (\tilde{m}^*_{ij})_{N\times N}$ is also one. Therefore, the summation of the coefficients of the $r$-th component is

$$\sum_i m^*_{ri} \sum_{s=1}^{4} \sum_j q^{(s)}_{ij} = \sum_i \tilde{m}^*_{ri} \left( \sum_j \tilde{w}_{ij} \right) = \sum_i \tilde{m}^*_{ri} = 1,$$

where $\mathbf{Q}_s = (q_{ij}^{(s)})_{N\times N}$ for $s = 1, 2, 3, 4$. As a result, we prove that the cell average $\bar{\mathbf{u}}$ is a convex combination of $\mathbf{f}$. Furthermore, we have $\bar{\mathbf{u}} \in [m, M]$. □

By Theorem 2.2, we get $\bar{\mathbf{u}} \in [m, M]$ for the fourth order implicit LDG scheme. Similarly, by repeating the above steps, we obtain the following Theorem 2.3 for $k = 1, 5, 7, 9$. According to the theorem, we can see that the higher order scheme we use, the smaller the lower bound of the CFL number is. Moreover, as long as we have $\bar{\mathbf{u}} \in [m, M]$, we can add the scaling MPP limiter to $u(x)$ and get our implicit high order maximum-principle-preserving numerical solutions.

**Theorem 2.3.** *For $k = 1, 5, 7, 9$, suppose $\mathbf{f}_\varrho \in [m, M]$ where $\varrho = 0, \cdots, k$ and the $j$-th component of the vector $\mathbf{f}_\varrho$ represents the function value of $f(x)$ at the $\varrho$-th LGL quadrature point in the $j$-th cell. Then as long as $\sigma \geq \sigma_{\min}^k$, we will have $\bar{\mathbf{u}} \in [m, M]$, where $\sigma_{\min}^k > 0$ is given in the following Table 2.1.*

| $k$ | *1* | *5* | *7* | *9* |
|---|---|---|---|---|
| $\sigma_{\min}^k$ | *0.0556* | *0.0379* | *0.0238* | *0.0157* |

TABLE 2.1. The lower bound $\sigma$ for cell average in $[m, M]$ when $k = 1, 5, 7, 9$.

**Proof:** For each $k$, the proof follows the same line as that in Theorem 2.2. We omit the tedious algebra here. □

2.1.2. *General theorem on the MPP property for any even order LDG schemes.* The previous section presents the results for $k = 1, 3, 5, 7, 9$. Here we present a general theorem on the MPP property for any odd number $k$. In this section, we only present the conclusions of the theorems, and defer the detailed proof to the appendix.

The "periodic tridiagonal" theorem of any even order schemes, similar to Theorem 2.1, is as follows:

**Theorem 2.4.** *Denote $\bar{\mathbf{u}} = (\bar{u}_1, \cdots, \bar{u}_N)^\top$ and $\mathbf{f}^\pm = (f^\pm_{1\mp\frac{1}{2}}, \cdots, f^\pm_{N\mp\frac{1}{2}})^\top$, then we have*

$$\mathbf{M}\bar{\mathbf{u}} = \sum_{\varrho=0}^{k} w_\varrho \mathbf{R}_\varrho \mathbf{f}^*_\varrho + \mathbf{T}_4 \mathbf{f}^+, \tag{2.28}$$

*where $\mathbf{M} = \mathbf{D}^{-1}(\mathbf{T}_1\mathbf{T}_3 - \mathbf{T}_2\mathbf{T}_4)$ and*

$$\mathbf{T}_1 = \mathbf{D}^\top\mathbf{D} + \sigma^{\frac{k+1}{2}}\frac{(2k+1)!}{k!}\mathbf{A}^\top\mathbf{D} + 4\sigma\mathbf{B}^\top\mathbf{B}, \quad \mathbf{T}_2 = 2\sigma^{\frac{k+1}{2}}\frac{(2k+1)!}{k!}\mathbf{B}^\top,$$
$$\mathbf{T}_3 = \mathbf{D} + \sigma^{\frac{k+1}{2}}\frac{(2k+1)!}{k!}\mathbf{A}, \quad \mathbf{T}_4 = 2\sigma\mathbf{A}\mathbf{B},$$

*in which*

$$\mathbf{A} = \mathbf{I} - \mathbf{G}^{N-1}, \quad \mathbf{A}^\top = \mathbf{I} - \mathbf{G}, \quad \mathbf{B} = -\gamma^*\mathbf{I} + \omega^*\mathbf{G},$$
$$\mathbf{B}^\top = -\gamma^*\mathbf{I} + \omega^*\mathbf{G}^{N-1}, \quad \mathbf{D} = (1+\xi)\mathbf{I} - \eta\mathbf{G}^{N-1}, \quad \mathbf{D}^\top = (1+\xi)\mathbf{I} - \eta\mathbf{G},$$

*and* $\mathbf{R}_\varrho = r_{1,\varrho}\mathbf{I} + r_{2,\varrho}\mathbf{G} + r_{3,\varrho}\mathbf{G}^{N-1}$ *for* $\varrho = 0, \cdots, k$, *with*

$$\begin{aligned} r_{1,\varrho} =& \frac{1}{2}\sum_{i=0}^{\frac{k-3}{2}}(4\sigma)^{i+1}\delta_1^{(2i+1)}(x_\varrho)(-1-\xi^*-\eta^*) + \frac{1}{2}\bigg((1+\xi)^2+\eta^2+\sigma^{\frac{k+1}{2}}\frac{(2k+1)!}{k!}(1+\xi+\eta) \\ &+ 4\sigma((\gamma^*)^2+(\omega^*)^2)\bigg) - \frac{1}{2}\bigg(4\sigma(\gamma^*+\omega^*)\bigg)\sum_{i=1}^{\frac{k-1}{2}}(4\sigma)^i\delta_1^{(2i)}(-x_\varrho), \end{aligned} \tag{2.29}$$

$$r_{2,\varrho} = \frac{1}{2}\sum_{i=0}^{\frac{k-3}{2}}(4\sigma)^{i+1}\delta_1^{(2i+1)}(x_\varrho)\eta^* + \frac{1}{2}\bigg(-(1+\xi)\eta^* + 4\sigma\gamma^*\omega^*\bigg) + \frac{1}{2}\sum_{i=1}^{\frac{k-1}{2}}(4\sigma)^i\delta_1^{(2i)}(-x_\varrho)(4\sigma\omega^*),$$

$$r_{3,\varrho} = \frac{1}{2}\bigg(\sum_{i=1}^{\frac{k-1}{2}}(4\sigma)^i\delta_1^{(2i-1)}(x_\varrho) - \eta\bigg)(1+\xi^*) + 2\sigma\gamma^*\bigg(\sum_{i=1}^{\frac{k-1}{2}}(4\sigma)^i\delta_1^{(2i)}(-x_\varrho) - \omega^*\bigg),$$

*where*

$$\xi = \sum_{i=0}^{\frac{k-3}{2}}(4\sigma)^{i+1}\delta_1^{(2i+1)}(1), \quad \eta = \sum_{i=0}^{\frac{k-3}{2}}(4\sigma)^{i+1}\delta_1^{(2i+1)}(-1), \quad \xi^* = \sum_{i=0}^{\frac{k-1}{2}}(4\sigma)^{i+1}\delta_1^{(2i+1)}(1),$$

$$\eta^* = \sum_{i=0}^{\frac{k-1}{2}}(4\sigma)^{i+1}\delta_1^{(2i+1)}(-1), \quad \omega^* = \sum_{i=0}^{\frac{k-1}{2}}(4\sigma)^i\delta_1^{(2i)}(1), \quad \gamma^* = \sum_{i=0}^{\frac{k-1}{2}}(4\sigma)^i\delta_1^{(2i)}(-1).$$

*Here* $\delta_1^{(i)}(x)$ *is the* $i$*-th derivative of* $\delta_1(x)$. *The* $j$*-th component of the vector* $\mathbf{f}_\varrho^* \in \mathcal{R}^N$ *for each* $\varrho = 0, \cdots, k$, *represents the function value of* $f(x)$ *at the* $\varrho$*-th LGL quadrature point in the* $j$*-th cell and* $\{w_\varrho\}_{\varrho=0}^k$ *are the corresponding LGL weights.*

**Proof:** The details of this proof are shown in Appendix B.1. □

The following theorem is similar to Theorem 2.2 and provides a sufficient condition to ensure the MPP property for the general even order LDG schemes.

**Theorem 2.5.** *Assume that* $k \geq 3$ *and* $\mathbf{f}_\varrho \in [m, M]$ *for* $\varrho = 0, \cdots, k$. *The equation* (2.28) *enjoys the following properties:*

*(i)* $\mathbf{M}$ *is an* $M$*-matrix when* $\sigma \geq \frac{1}{2}$.

*(ii) When* $\sigma$ *satisfies the following condition:*

$$\sigma \geq \frac{1}{4} + \max_{\varrho=\{1,\cdots,k-1\}}\left(\frac{1}{1-|x_\varrho|}\left(\frac{4}{9}+\frac{2k+2}{2k+1}\right)\right), \tag{2.30}$$

*where* $\{x_\varrho\}_{\varrho=0}^k$ *are the LGL points in the reference interval, we have the coefficient matrices* $\mathbf{R}_\varrho$ *are positive when* $\varrho = 1, \cdots, k$ *and* $\mathbf{R}_0 + \mathbf{T}_4$ *is positive.*

*(iii) All row sums of matrix* $\mathbf{M}$ *are equal and equal to all row sums of matrix* $\sum_{\varrho=0}^k \mathbf{R}_\varrho + \mathbf{T}_4$. *Furthermore, we can conclude* $\bar{\mathbf{u}} \in [m, M]$. *Indeed, the above properties indicate the cell average vector is a convex combination of point value vectors.*

**Proof:** The proof of (i) is to check the definition of $M$-matrices and the details can be found in Appendix B.2. The proof of (ii) is the most difficult. The main technique of

this proof is to utilize the highest derivative to control the others. But doing so will make some inequalities less sharp, leading to less sharp lower bounds. The details can be found in Appendix B.3. The property (iii) can be checked directly and we omit the proof for simplicity. □

The advantage of Theorem 2.5 is that it provides a lower bound of the CFL number for any odd number $k \geq 3$, as shown in Table 2.2. It should be noted that the lower bound in Theorem 2.5 is not sharp because some inequalities that hold for all $k$ are not tight. We will explore how to make the lower bound sharper in our future work. We should also note that the lower bound in Theorem 2.2 and Theorem 2.3, even though not sharp, is not too far from the sharp bound, as verified by numerical experiments whose results are shown in Table 2.3.

| $k$ | $\sigma_{\min}$ by Theorem 2.2 and Theorem 2.3 | $\sigma_{\min}$ by general Theorem 2.5 |
|---|---|---|
| 1 | 0.0556 | - |
| 3 | 0.0544 | 3.1215 |
| 5 | 0.0379 | 6.7850 |
| 7 | 0.0238 | 12.0317 |
| 9 | 0.0157 | 18.8551 |

TABLE 2.2. Comparison of the lower bounds of the CFL numbers in Theorem 2.2, Theorem 2.3 and Theorem 2.5.

2.1.3. *The MPP property of the third order LDG scheme.* Here we consider the MPP property of the LDG scheme when $k = 2$. When $k$ is even, the proof is different from the proofs when $k$ is odd because of the different representations of the cell averages. For this third order scheme, the property $(ii)$ in Theorem 2.2 or Theorem 2.5 does not hold. To overcome this difficulty, we follow the approach in [6] to get the inverse of the matrix on the left of the equality explicitly and derive the CFL condition for the MPP property. Similarly with Theorem 2.2, by using the $\delta_y(x)$ function repeatedly, we have the following "periodic tridiagonal" theorem for the $P^2$ scheme.

**Theorem 2.6.** *Denote* $\bar{\mathbf{u}} = (\bar{u}_1, \cdots, \bar{u}_N)^\top$, $\mathbf{f}^+ = (f^+_{\frac{1}{2}}, \cdots, f^+_{N-\frac{1}{2}})^\top$, $\mathbf{f}^- = (f^-_{\frac{3}{2}}, \cdots, f^-_{N+\frac{1}{2}})^\top$ *and* $\mathbf{f}^0 = (f_1, \cdots, f_N)^\top$ *where* $f_i$ *is the value of function* $f$ *at the center of the* $i$*-th cell. Then we have*

$$\mathbf{M}\bar{\mathbf{u}} = \mathbf{Q}_1\mathbf{f}^+ + \mathbf{Q}_2\mathbf{f}^0 + \mathbf{Q}_3\mathbf{f}^-, \tag{2.31}$$

*where* $\mathbf{M} = \mathbf{T}_1 - 60\sigma\mathbf{T}_4$ *and*

$$\mathbf{Q}_1 = \frac{1}{6}\mathbf{T}_1 + \mathbf{T}_4 + 4\sigma\mathbf{T}_3^{-1}(-3\mathbf{T}_4\mathbf{B}^\top + \mathbf{T}_1\mathbf{A}),$$
$$\mathbf{Q}_2 = \frac{2}{3}\mathbf{T}_1 - 4\sigma\mathbf{T}_3^{-1}(-3\mathbf{T}_4\mathbf{B}^\top + \mathbf{T}_1\mathbf{A}),$$
$$\mathbf{Q}_3 = \frac{1}{6}\mathbf{T}_1 - 6\sigma\mathbf{T}_3^{-1}(-3\mathbf{T}_4\mathbf{B}^\top + \mathbf{T}_1\mathbf{A}),$$

*in which*

$$\mathbf{A}=\mathbf{I}-\mathbf{G}^{N-1},\ \mathbf{B}=-\mathbf{I}+3\mathbf{G},\ \mathbf{D}=(1+36\sigma)\mathbf{I}+24\sigma\mathbf{G}^{N-1}, \tag{2.32}$$

*and*

$$\begin{aligned}
\mathbf{T}_1&=\mathbf{D}\mathbf{D}^\top+9\sigma\mathbf{B}^\top\mathbf{B}-180\sigma^2\mathbf{A}^\top\mathbf{B}^\top\\
&=(1+162\sigma+2592\sigma^2)\mathbf{I}+(684\sigma^2-3\sigma)\mathbf{G}+(-3\sigma+324\sigma^2)\mathbf{G}^{N-1},\\
\mathbf{T}_2&=60\sigma\mathbf{D}=(60\sigma+2160\sigma^2)\mathbf{I}+1440\sigma^2\mathbf{G}^{N-1},\\
\mathbf{T}_3&=\mathbf{D}=(1+36\sigma)\mathbf{I}+24\sigma\mathbf{G}^{N-1},\\
\mathbf{T}_4&=3\sigma\mathbf{A}\mathbf{B}-60\sigma^2\mathbf{A}\mathbf{A}^\top=(-12\sigma-120\sigma^2)\mathbf{I}+(9\sigma+60\sigma^2)\mathbf{G}+(3\sigma+60\sigma^2)\mathbf{G}^{N-1}.
\end{aligned} \tag{2.33}$$

**Proof:** This proof is similar to that of Theorem 2.1 and Theorem 2.4, hence we omit the details to save space. □

Notice that

$$\begin{aligned}
\mathbf{T}_1-60\sigma\mathbf{T}_4=&(1+162\sigma+3312\sigma^2+7200\sigma^3)\mathbf{I}\\
&+(-3600\sigma^3+144\sigma^2-3\sigma)\mathbf{G}+(-3600\sigma^3+144\sigma^2-3\sigma)\mathbf{G}^{N-1}.
\end{aligned} \tag{2.34}$$

It is clear that if $\sigma>0.056$, then $\mathbf{M}$ is an $M$-matrix. Unfortunately, the coefficient matrices $\mathbf{Q}_i$ $(i=1,2,3)$ are not all positive matrices. However, we have the following lemma (see [6]) for getting the inverse of the matrix $\mathbf{M}$. Then we can invert the left matrix to the right, and prove the new coefficient matrices of $\mathbf{f}^\varrho$, where $\varrho=\pm,0$, are all positive. The lemma is as below.

**Lemma 2.1.** *Assume a matrix $\mathbf{E}\in\mathbb{R}^{N\times N}$ has the following form*

$$\mathbf{E}=\zeta_1\mathbf{I}+\zeta_2(\mathbf{G}+\mathbf{G}^{N-1})+\zeta_3(\sum_{k=2}^{N-2}\mathbf{G}^k),$$

*where $\zeta_i\in\mathbb{R},\ i=1,2,3$. If $\mathbf{E}$ is invertible and*

$$(\zeta_1+2\zeta_2+(N-3)\zeta_3)\prod_{i_1=1}^{\lceil\frac{N-1}{2}\rceil}\big[\zeta_1-\zeta_3+2(\zeta_2-\zeta_3)\cos\big(\frac{2\pi i_1}{N}\big)\big]\neq 0, \tag{2.35}$$

*then we have*

$$\mathbf{E}^{-1}=C(\zeta_1,\zeta_2,\zeta_3,\zeta_3,\cdots,\zeta_3,\zeta_3,\zeta_2)^{-1}=C(g_1(\zeta_1,\zeta_2,\zeta_3),\cdots,g_N(\zeta_1,\zeta_2,\zeta_3)), \tag{2.36}$$

*where the notation $C(a_1,\cdots,a_N)$ is a representation of circulant matrices, akin to* (2.3). *If $\zeta_1\neq 3\zeta_3-2\zeta_2$, then $g_{i_1}(\zeta_1,\zeta_2,\zeta_3)$ has the form:*

$$g_{i_1}(\zeta_1,\zeta_2,\zeta_3)=\frac{U^{i_1-2}(z)+U^{N-i_1}(z)}{2(\zeta_3-\zeta_2)[T^N(z)-1]}-\frac{\zeta_3}{(\zeta_1+2\zeta_2-3\zeta_3)(\zeta_1+2\zeta_2+(N-3)\zeta_3)}, \tag{2.37}$$

*where $z=\frac{\zeta_3-\zeta_1}{2(\zeta_2-\zeta_3)}$, $i_1=1,\cdots,N$, and $T^{i_1},U^{i_1}$ are the first and the second classes of Chebyshev polynomials respectively.*

**Proof:** For the proof of this lemma, we recommend the interested readers to [6]. □

By the lemma above, we can get the following MPP theorem for the implicit $P^2$ LDG scheme.

**Theorem 2.7.** *The $P^2$ scheme has the following properties:*

*(i) $\mathbf{M}^{-1}\mathbf{Q}_i$ are positive when $\sigma > 0.162$ for $i = 1, 2, 3$.*

*(ii) All row sums of the matrix $\mathbf{M}$ are the same, and are equal to all the row sums of the matrix $\sum_{i=1}^{3} \mathbf{Q}_i$.*

**Proof:** The key for the proof is to get a concise expression of $\mathbf{M}^{-1}\mathbf{T}_3(-3\mathbf{T}_4\mathbf{B}^\top + \mathbf{T}_1\mathbf{A})$. We leave the details to Appendix B.4. □

The result in Theorem 2.7 is similar to Theorem 2.2. Numerical experiments show that the lower bounds in these theorems are not far from being sharp. The specific numerical experiments are as follows. We take the initial conditions $u_0^{(k)}$ $(k = 1, 2, 3)$ for the implicit $P^k$ $(k = 1, 2, 3)$ schemes to be as below:

$$u_0^{(1)}(x) = \begin{cases} 1, \ if \ x = 0, \\ 0, \ else, \end{cases} \quad u_0^{(2)}(x) = u_0^{(1)}(x), \quad u_0^{(3)}(x) = \begin{cases} 1, \ if \ x = \dfrac{h}{2}\left(1 - \dfrac{1}{\sqrt{5}}\right), \\ 0, \ else. \end{cases}$$

Based on the above initial conditions, we can numerically test the lower bounds for the MPP property and denote those lower bounds as $\sigma_{\min}^N$. Also for the purpose of comparison, we denote those theoretical lower bounds in Theorem 2.2, Theorem 2.3 and Theorem 2.7 as $\sigma_{\min}^T$. Note that the lower bounds of $\sigma_{\min}^N$ is a necessary condition and $\sigma_{\min}^T$ is a sufficient condition for the MPP property. Hence we conclude that the sharp lower bound of $\sigma$ lies in $[\sigma_{\min}^N, \sigma_{\min}^T]$. The table below shows the comparison between theoretical and numerical experimental results, and it can be seen that the lower bounds in our proof are not far from being sharp.

| $k$ | Theoretical $\sigma_{\min}^T$ | Numerical $\sigma_{\min}^N$ | $\min_i \bar{u}_i^1$ | $\sigma_{\min}^N - 0.0001$ | $\min_i \bar{u}_i^1$ |
|---|---|---|---|---|---|
| 1 | 0.0556 | 0.0556 | 5E-16 | 0.0555 | -1E-06 |
| 2 | 0.1620 | 0.1078 | 1E-08 | 0.1077 | -4E-06 |
| 3 | 0.0544 | 0.0393 | 2E-12 | 0.0392 | -2E-05 |

TABLE 2.3. Comparison of the lower bounds of the CFL numbers in Theorem 2.2, Theorem 2.3, Theorem 2.7

and the numerically observed lower bounds.

2.2. **The MPP property of the second order scheme in 2D.** In this subsection, we consider the implicit second order LDG scheme for solving linear diffusion equations in 2D, which can be viewed as the tensor product generalization of the previous subsection. We will use $(x, y)$ to replace $\mathbf{x}$ in (1.1). The proof in this section is similar to the 1D case but the details are much more tedious. Firstly, we consider the following equation:

$$u_t = u_{xx} + u_{yy}, \ (x, y) \in \Omega, \tag{2.38}$$

with periodic boundary condition. Divide $\Omega = [0, 1] \times [0, 1]$ into $N_x \times N_y$ uniform cells $\Omega = \bigcup_{i,j} I_{ij}$ where $I_{ij} = [x_{i-\frac{1}{2}}, x_{i+\frac{1}{2}}] \times [y_{j-\frac{1}{2}}, y_{j+\frac{1}{2}}]$. Define the mesh step sizes as $h_x = x_{i+\frac{1}{2}} - x_{i-\frac{1}{2}}$ and $h_y = y_{j+\frac{1}{2}} - y_{j-\frac{1}{2}}$ and the 2D reference cell as $\hat{I} = [-1, 1] \times [-1, 1]$. We choose the function space as $V_h = \{u_h(x, y) \in L^2(\Omega) : u_h|_{I_{ij}} \in Q^1(I_{ij})\}$ where $Q^1(\hat{I})$

is the polynomial space of the tensor product of $P^1([-1,1])$ in 1D. Then the LDG scheme is to find $u, q^1, q^2 \in V_h$ such that, for all $v, w, \zeta \in V_h$, we have

(2.39a)

$$(\frac{u-f}{\tau}, v)_{\Omega_{ij}} = -(q^1, v_x)_{\Omega_{ij}} - (q^2, v_y)_{\Omega_{ij}} + \int_{y_{j-\frac{1}{2}}}^{y_{j+\frac{1}{2}}} \left( \widehat{q^1}(x_{i+\frac{1}{2}}, y) v(x^-_{i+\frac{1}{2}}, y) - \right.$$

$$\left. \widehat{q^1}(x_{i-\frac{1}{2}}, y) v(x^+_{i-\frac{1}{2}}, y) \right) dy + \int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} \left( \widehat{q^2}(x, y_{j+\frac{1}{2}}) v(x, y^-_{j+\frac{1}{2}}) - \widehat{q^2}(x, y_{j-\frac{1}{2}}) v(x, y^+_{j-\frac{1}{2}}) \right) dx,$$

(2.39b)

$$(q^1, w)_{\Omega_{ij}} = -(u, w_x)_{\Omega_{ij}} + \int_{y_{j-\frac{1}{2}}}^{y_{j+\frac{1}{2}}} \left( \widehat{u}(x_{i+\frac{1}{2}}, y) w(x^-_{i+\frac{1}{2}}, y) - \widehat{u}(x_{i-\frac{1}{2}}, y) w(x^+_{i-\frac{1}{2}}, y) \right) dy,$$

(2.39c)

$$(q^2, \zeta)_{\Omega_{ij}} = -(u, \zeta_y)_{\Omega_{ij}} + \int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} \left( \widehat{u}(x, y_{j+\frac{1}{2}}) \zeta(x, y^-_{j+\frac{1}{2}}) - \widehat{u}(x, y_{j-\frac{1}{2}}) \zeta(x, y^+_{j-\frac{1}{2}}) \right) dx.$$

We take the numerical fluxes as the alternating fluxes, namely $\widehat{q^i} = q^{i,-} (i = 1, 2)$ which are the left or the lower limits of the cell boundaries and $\hat{u} = u^+$ which are the right or the upper limits of the cell boundaries. We define the basis functions on $\hat{I}$ as follows:

$$\begin{aligned} l_1(\alpha, \beta) &= \phi_0(\alpha)\phi_0(\beta), \qquad l_2(\alpha, \beta) = \phi_1(\alpha)\phi_0(\beta), \\ l_3(\alpha, \beta) &= \phi_0(\alpha)\phi_1(\beta), \qquad l_4(\alpha, \beta) = \phi_1(\alpha)\phi_1(\beta), \end{aligned} \tag{2.40}$$

where

$$\phi_0(\alpha) = \frac{1-\alpha}{2}, \quad \phi_1(\alpha) = \frac{1+\alpha}{2}, \quad \alpha \in [-1, 1], \quad \beta \in [-1, 1].$$

**Theorem 2.8.** *By* (2.39)*, we have the following matrix form representation of the* $Q^1$ *LDG scheme:*

$$\mathbf{F}\bar{\mathbf{u}} = \mathbf{W}_1\mathbf{f}_1 + \mathbf{W}_2\mathbf{f}_2 + \mathbf{W}_3\mathbf{f}_3 + \mathbf{W}_4\mathbf{f}_4, \tag{2.41}$$

*where the vectors* $\mathbf{f}_\varrho \in \mathcal{R}^{N_x N_y}$ *are coefficient vectors of* $l_\varrho(\alpha, \beta)$ *for* $\varrho = 1, 2, 3, 4$*, respectively. The definitions of* $\mathbf{F}$ *and* $\mathbf{W}_\varrho$ $(\varrho = 1, 2, 3, 4)$ *are given in Appendix B.5.*

**Proof:** Here we still use $\delta^k(x)$ to finish the proof which is similar with the 1D case. The proof is given in Appendix B.6. □

**Theorem 2.9.** *Suppose* $\mathbf{f}_\varrho \in [m, M]$ *for* $\varrho = 1, 2, 3, 4$*. If we take* $h = h_x = h_y$*, then the* $Q^1$ *LDG scheme has the following properties:*

*(i) The matrix* $\mathbf{F}$ *is an* $M$*-matrix when* $\sigma = \frac{\tau}{h^2} \geq 1$*.*

*(ii) The matrices* $\mathbf{W}_\varrho$ $(\varrho = 1, 2, 3, 4)$ *are all positive when* $\sigma > 0$*.*

*(iii) All the row sums of the matrix* $\mathbf{F}$ *are equal and they are equal to all the row sums of the matrix* $\sum_{\varrho=1}^{4} \mathbf{W}_\varrho$*.*

*Then we get* $\bar{\mathbf{u}} \in [m, M]$ *when* $\sigma \geq 1$*.*

**Proof:** The proof is an easy generalization of the 1D case and is given in Appendices B.7 and B.8. □

2.3. **The scaling limiter.** Once we prove the cell averages are bounded by $m$ and $M$, where we recall that $m = \min_{\mathbf{x}} u_0(\mathbf{x})$ and $M = \max_{\mathbf{x}} u_0(\mathbf{x})$, we can use the following scaling limiter to modify the DG polynomial and make the modified polynomial $\tilde{u}^{n+1}(\mathbf{x})$ bounded. Here we take the one dimensional scheme as an example and use $x$ instead of $\mathbf{x}$. The usage of the scaling MPP limiter is as follows.

STEP 1: Denote $\bar{u}_j^n$ as the cell average of $u^n(x)$ on the $j$-th cell and suppose that $m \leq u^n(x) \leq M$ for all $x \in \cup_j S_j$ . Recall that $S_j$ is the set of LGL quadrature points for $\Omega_j$. Then we obtain $u^{n+1}(x)$ by the implicit LDG scheme.

STEP 2: We modify the polynomial $u^{n+1}(x)$ in the cell $I_j$ as follows:

$$\tilde{u}^{n+1}(x) = \theta(u^n(x) - \bar{u}_j^n) + \bar{u}_j^n,$$

with

$$\theta = \min\left\{\left|\frac{M - \bar{u}_j^n}{M_j - \bar{u}_j^n}\right|, \left|\frac{m - \bar{u}_j^n}{m_j - \bar{u}_j^n}\right|, 1\right\}.$$

Here $M_j = \max_{x\in S_j} u^{n+1}(x),\ m_j = \min_{x\in S_j} u^{n+1}(x)$. Then the modified function $\tilde{u}^{n+1}(x)$ is our numerical solution at the time level $t^{n+1}$.

## 3. Second order maximum-principle-preserving implicit LDG scheme for linear convection-diffusion equations

In this section, we consider the MPP property of the implicit LDG scheme without any limiter for solving linear convection-diffusion equations with constant coefficients and periodic boundary condition in 1D, which means that

$$\Phi(u) = -cu, \quad \kappa = a, \tag{3.1}$$

and $\Omega = [0, 1]$ in (1.1), where $a$ and $c$ are positive constants. Here we can consider nonuniform meshes. The implicit LDG scheme is: find $u, q \in V_h$ such that, for all $v, w \in V_h$, we have

(3.2)

$$c\tau\left((u, v_x)_j + (u^+_{j+\frac{1}{2}} v^-_{j+\frac{1}{2}} - u^+_{j-\frac{1}{2}} v^+_{j-\frac{1}{2}})\right) - \sqrt{a}\tau\left(q^-_{j+\frac{1}{2}} v^-_{j+\frac{1}{2}} - q^-_{j-\frac{1}{2}} v^+_{j-\frac{1}{2}} - (q, v_x)_j\right) = (u - f, v)_j,$$

$$(q, w)_j - \sqrt{a}\left(u^+_{j+\frac{1}{2}} w^-_{j+\frac{1}{2}} - u^+_{j-\frac{1}{2}} w^+_{j-\frac{1}{2}} - (u, w_x)_j\right) = 0,$$

where $\tau$ is the time step and $f$ represents $u^n$ which is the numerical solution at the current time level $t^n$, and $u$ denotes the numerical solution $u^{n+1}$ at the next time level $t^{n+1}$. We again take $\hat{I} = [-1, 1]$ as our reference cell and choose the LGL type basis for analysis.

Through easy algebraic calculation, we can get the following matrix form of the implicit LDG scheme:

$$\mathcal{L}\mathbf{u} = \mathbf{M}\mathbf{f}, \tag{3.3}$$

where

$$\mathcal{L} = \mathbf{M} - \tau\left(4a\mathbf{H}^{-1}\mathbf{D_q}\mathbf{H}^{-1}\mathbf{M}^{-1}\mathbf{D_u} + 2c\mathbf{H}^{-1}\mathbf{D_u}\right).$$

The definitions of the block matrices $\mathbf{M}, \mathbf{H}, \mathbf{D_u}, \mathbf{D_q} \in \mathcal{R}^{2N\times 2N}$ are as follows. When we say a block matrix belongs to $\mathcal{R}^{2N\times 2N}$, it means that it can be viewed as a matrix in $\mathcal{R}^{N\times N}$ whose elements are all in $\mathcal{R}^{2\times 2}$. Denote $\mathbf{M}_{ij}$ as the $(i,j)$-th block of the matrix $\mathbf{M}$ and define the indexes pair $(N, N+1) = (N, 1)$ and $(0, 1) = (N, 1)$ which means that the indexes are in the sense of mod $N$. Then, for each $i = 1, \cdots, N$, we denote

$$\mathbf{M}_{ii} = \mathbf{M_a}, \quad \mathbf{H}_{ii} = h_i \mathbf{I}_2, \quad \mathbf{I}_2 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \quad h_i = x_{i+\frac{1}{2}} - x_{i-\frac{1}{2}},$$

$$(\mathbf{D_q})_{ii} = -\mathbf{D} + \mathbf{B}^1, \quad (\mathbf{D_q})_{i,i-1} = -\mathbf{B}^2, \quad (\mathbf{D_u})_{ii} = -\mathbf{D} - \mathbf{B}^4, \quad (\mathbf{D_u})_{i,i+1} = \mathbf{B}^3,$$

and

$$\mathbf{M_a} = \begin{pmatrix} \frac{2}{3} & \frac{1}{3} \\ \frac{1}{3} & \frac{2}{3} \end{pmatrix}, \quad \mathbf{D} = \begin{pmatrix} -\frac{1}{2} & -\frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} \end{pmatrix}, \quad \mathbf{B}^1 = \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}, \tag{3.4}$$
$$\mathbf{B}^2 = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}, \quad \mathbf{B}^3 = \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix}, \quad \mathbf{B}^4 = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}.$$

Note that all of the matrices are block matrices, therefore we have

$$(2\mathbf{H}^{-1}\mathbf{D_u})_{ii} = \begin{pmatrix} -\frac{1}{h_i} & \frac{1}{h_i} \\ -\frac{1}{h_i} & -\frac{1}{h_i} \end{pmatrix}, \quad (2\mathbf{H}^{-1}\mathbf{D_u})_{i,i+1} = \begin{pmatrix} 0 & 0 \\ \frac{2}{h_i} & 0 \end{pmatrix}, \tag{3.5}$$

and

$$(4\mathbf{H}^{-1}\mathbf{D_q}\mathbf{H}^{-1}\mathbf{M}^{-1}\mathbf{D_u})_{i,i-1} = \begin{pmatrix} \frac{2}{h_i h_{i-1}} & \frac{6}{h_i h_{i-1}} \\ 0 & 0 \end{pmatrix}, \tag{3.6a}$$

$$(4\mathbf{H}^{-1}\mathbf{D_q}\mathbf{H}^{-1}\mathbf{M}^{-1}\mathbf{D_u})_{i,i} = \begin{pmatrix} -\frac{2}{h_i^2} - \frac{8}{h_i h_{i-1}} & 0 \\ 0 & -\frac{6}{h_i^2} \end{pmatrix}, \tag{3.6b}$$

$$(4\mathbf{H}^{-1}\mathbf{D_q}\mathbf{H}^{-1}\mathbf{M}^{-1}\mathbf{D_u})_{i,i+1} = \begin{pmatrix} \frac{2}{h_i^2} & 0 \\ \frac{6}{h_i^2} & 0 \end{pmatrix}. \tag{3.6c}$$

The other elements of these matrices without being mentioned are all zero. It is clear that the off-diagonal block matrices are non-positive, but all the diagonal block matrices are strictly positive (i.e., each element is larger than zero). Therefore we cannot use the properties of $M$-matrices directly. However, thanks to the work of Bouchon ([5], Theorem 2.5), we can view the positive off-diagonal elements as perturbations of an $M$-matrix to get the inverse-positive property of the matrices on the left of (3.3). Let us start by some definitions.

**Definition 3.1.** *A matrix $\mathbf{A} = (a_{ij}) \in \mathcal{R}^{N\times N}$ is irreducible, if for all $i \neq j$, there is always a positive number $r \in \mathcal{N}$, where $\mathcal{N}$ is the set of all positive integers, and a chain $i = i_0, \cdots, i_r = j$ s.t. $a_{i_{l-1},i_l} \neq 0$ for all $l = 1, \cdots, r$.*

**Definition 3.2.** *An irreducibly diagonally dominant matrix $\mathbf{A} \in \mathcal{R}^{N\times N}$ means that $\mathbf{A}$ is both irreducible and diagonally dominant, with strict dominance on at least one row (i.e., for any $i \in \{1, \cdots, N\}$, we have $|a_{ii}| \geq \sum_{j\neq i} |a_{ij}|$ and $|a_{ii}| > \sum_{j\neq i} |a_{ij}|$ holds for at least one $i$).*

Assume $\mathbf{A} = (a_{ij})$ and $\tilde{\mathbf{A}} = (\tilde{a}_{ij})$ are two matrices in $\mathcal{R}^{N\times N}$, we define the following sets:

$$\forall (i,j) \in \{1,\cdots,N\}^2,\ C_{i,j} = \Big\{ r \in \mathcal{N} \Big|\ \exists (i_0,\cdots,i_r) \in \mathcal{N}^{r+1},$$
$$s.t.\ i_0 = i,\ i_r = j, a_{i_{l-1}i_l} \neq 0, \forall l \in \{1,\cdots,r\} \Big\}.$$

Then we define the following variables:

$$b(\mathbf{A}) = \min_{i=1,\cdots,N}(|a_{ii}|), \quad \zeta(\mathbf{A}) = \max_{\substack{1\le i,j\le N \\ s.t.\ a_{ij}\neq 0}} \frac{|a_{ii}|}{|a_{ij}|}, \quad B = \max_{\substack{i,j\in\{1,\cdots,N\} \\ s.t.\ \tilde{a}_{ij}\neq 0}} (d(i,j)),$$

where $d(i,j) = \min\{r \in C_{i,j}\}$, $d(i,i) = 0,\ \forall i$ and $d(i,j) = \infty$ if the set $C_{i,j}$ is empty. Furthermore, we have the following Bouchon's lemma.

**Lemma 3.1.** *Let $\mathbf{A} = (a_{ij})$ and $\tilde{\mathbf{A}} = (\tilde{a}_{ij})$ be two matrices in $\mathcal{R}^{N\times N}$ with the following properties:*

*(1) $\mathbf{A}$ is an irreducibly diagonally dominant M-matrix,*

*(2) $\tilde{\mathbf{A}}\mathbf{1} \geq 0$.*

*Then $||\tilde{\mathbf{A}}||_\infty < Cb(\mathbf{A})$ with*

$$C = \frac{1}{(\zeta(\mathbf{A}))^B Be}, \tag{3.7}$$

*where e is the natural constant and approximately equals to 2.7182818 and $\mathbf{1} = (1,\cdots,1)^\top$. When we say a vector is nonnegative, it means that each component in the vector is nonnegative.*

**Proof:** This proof is from Bouchon. We refer the interested readers to [5]. □

As long as we can verify the conditions in the previous lemma to prove $\mathcal{L}$ is inverse-positive then we can ensure the scheme is lower-bound preserving. Also, going back to the equation (3.3), the sum of each row of the matrix $\mathcal{L}$ is equal to that of the matrix $\mathbf{M}$. Hence, once we get the inverse positivity of $\mathcal{L}$, then the MPP property can be derived directly. The following theorem provides a sufficient condition to ensure $\mathcal{L}$ is inverse positive.

**Theorem 3.1.** *Assume $\frac{c\tau}{h_i} < \frac{1}{3}$ and $\frac{h_i}{h_{i-1}} \geq \frac{1}{16}$ for $i = 1,\cdots,N$, where $h_0 = h_N$. Define $\sigma_i = \frac{\tau}{h_i^2}$, $h_{\max} = \max_i h_i$ and $\sigma_{\min} = \min_i \sigma_i$. If $\sigma_i = \tau/h_i^2$ and the mesh decomposition satisfies the following inequality for each i,*

$$\frac{4e}{3}\max\left\{1, \left(\frac{h_{i-1}}{h_i}\right)^2\right\}\left(\frac{1}{2a\sigma_i} + 1 + 4\frac{h_i}{h_{i-1}}\right)^2 < \frac{2}{3} + 6a\sigma_{\min} + c\frac{\tau}{h_{\max}}.$$

*Then $\mathcal{L}$ is inverse-positive. Furthermore, the scheme (3.2) satisfies the MPP property.*

**Proof:** The proof is a direct application of Bouchon's theorem. The details are shown in Appendix C.1. □

**Lemma 3.2.** *If we use the lumped mass, then $\mathcal{L}$ will be an M-matrix and the scheme is MPP directly.*

**Proof:** If we use the lumped mass, then $\mathbf{M_a}$ in (3.4) becomes the identity matrix

$$\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}.$$

The definition of $M$-matrices can then be easily checked and hence we can prove this lemma. □

**Remark**: In particular, if we take a uniform mesh, the condition in Theorem 3.1 becomes

$$(6a + ch)\sigma^3 + \left(\frac{2 - 100e}{3}\right)\sigma^2 - \frac{20e}{3a}\sigma - \frac{e}{3a^2} > 0.$$

Furthermore, notice that $c > 0$, then $a\sigma \geq 15.19$ would be a sufficient condition to ensure the MPP property.

**Remark**: The lower bound proved in Theorem 3.1 is far from sharp. In fact, for linear equations, we can explicitly write out the matrix elements and numerically invert them for a fixed mesh (taken here as 10 cells), thereby obtaining a lower bound that satisfies the conditions for any initial condition. For the convection diffusion equation, the lower bound obtained through the above procedure and the comparison with the lower bound obtained from the theorem are shown in Table 3.1 below.

| | $\sigma_{\min}$ in numerical test | | | $\sigma_{\min}$ in proof | |
|---|---|---|---|---|---|
| | $\times$ | $\circ$ | $\surd$ | $\bigcirc$ | $\surd$ |
| $P^1$ | $\leq 0.05$ | $\geq 0.06$ | $\geq 0.40$ | $\geq 0.06$ | $\geq 15.19$ |
| $P^2$ | $\leq 0.10$ | $\geq 0.11$ | $\geq 0.21$ | $\geq 0.17$ | - |
| $P^3$ | $\leq 0.01$ | $\geq 0.02$ | $\geq 0.09$ | $\geq 0.06$ | - |

TABLE 3.1. Comparison of lower bounds of the theoretical proof and the numerical validation for the convection diffusion equation (3.1). "$\times$" means $\sigma_{\min}$ for producing negative averages. "$\circ$" means $\sigma_{\min}$ for cell averages with the MPP property. "$\surd$" means $\sigma_{\min}$ for MPP point values. "$\bigcirc$" means $\sigma_{\min}$ for cell averages with the MPP property for the pure diffusion equation (2.1) which is proved in Theorem 2.2, Theorem 2.3 and Theorem 2.7.

## 4. Numerical tests

In this section, we present results of numerical experiments. First, we verify the high-order spatial accuracy of the scheme with the MPP limiter by testing it on both linear and nonlinear problems with time-independent exact solutions (to avoid the numerical errors being dominated by the first order temporal error from backward Euler) and both in 1D and 2D. Next, we use the schemes both with and without the limiter to solve the degenerate equations such as the porous medium equation and the Buckley-Leverett equation. Numerical results indicate that for nonlinear equations, the scheme without the limiter will produce negative values, while the scheme with the MPP limiter always satisfies the MPP property. In the following, when we say "the limiter", it means that "the MPP limiter".

4.1. **1D accuracy test.** We present the numerical results of our scheme with the limiter for solving 1D linear and nonlinear equations which have smooth solutions to validate accuracy. In these experiments, we take $\Omega = [0, 1]$.

4.1.1. *Linear diffusion equation.* Here, we consider the equation with $\Phi = 0, \kappa = 1$ in (1.1) with periodic boundary condition and a source term $s(x) = \sin(8\pi x)$ at the right-hand side of the equation. This problem has steady solution $u = \sin(8\pi x)/(8\pi)^2$. We take $m = -\frac{1}{(8\pi)^2}$ and $M = \frac{1}{(8\pi)^2}$ in the limiter. The ending time $T_{end} = 0.2$ and the results are shown in Table 4.1. We can clearly see that the limiter does not affect the designed order of accuracy. The notation "MPP(%)" represents the percentage of cells that utilize the limiter among all cells.

| | without the limiter | | | | | with the limiter | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| $P^1$ | mesh | $L^1$ error | order | $L^\infty$ error | order | $L^1$ error | order | $L^\infty$ error | order | MPP (%) |
| | 10 | 8.95E-04 | | 1.38E-03 | | 8.95E-04 | | 1.14E-03 | | 60.00 |
| | 20 | 2.57E-04 | 1.80 | 3.97E-04 | 1.80 | 2.57E-04 | 1.80 | 3.45E-04 | 1.73 | 40.00 |
| | 40 | 6.63E-05 | 1.95 | 1.02E-04 | 1.95 | 6.63E-05 | 1.95 | 7.75E-05 | 2.15 | 40.00 |
| | 80 | 1.66E-05 | 2.00 | 2.58E-05 | 1.99 | 1.66E-05 | 2.00 | 2.58E-05 | 1.59 | 10.00 |
| | 160 | 4.14E-06 | 2.00 | 6.50E-06 | 1.99 | 4.15E-06 | 2.00 | 6.50E-06 | 1.99 | 5.00 |
| $P^2$ | mesh | $L^1$ error | order | $L^\infty$ error | order | $L^1$ error | order | $L^\infty$ error | order | MPP (%) |
| | 10 | 1.54E-04 | | 3.56E-04 | | 1.59E-04 | | 3.44E-04 | | 40.00 |
| | 20 | 2.17E-05 | 2.82 | 5.03E-05 | 2.82 | 2.17E-05 | 2.87 | 5.03E-05 | 2.77 | 0.00 |
| | 40 | 2.69E-06 | 3.01 | 6.29E-06 | 3.00 | 2.69E-06 | 3.01 | 6.29E-06 | 3.00 | 0.00 |
| | 80 | 3.48E-07 | 2.95 | 8.10E-07 | 2.96 | 3.54E-07 | 2.93 | 8.10E-07 | 2.96 | 10.00 |
| | 160 | 4.34E-08 | 3.00 | 1.02E-07 | 2.99 | 4.36E-08 | 3.02 | 1.02E-07 | 2.99 | 5.00 |
| $P^3$ | mesh | $L^1$ error | order | $L^\infty$ error | order | $L^1$ error | order | $L^\infty$ error | order | MPP (%) |
| | 10 | 2.31E-05 | | 6.76E-05 | | 2.31E-05 | | 6.76E-05 | | 0.00 |
| | 20 | 1.54E-06 | 3.90 | 4.47E-06 | 3.92 | 1.46E-06 | 3.98 | 4.76E-06 | 3.83 | 20.00 |
| | 40 | 1.02E-07 | 3.92 | 2.92E-07 | 3.94 | 1.02E-07 | 3.84 | 2.92E-07 | 4.03 | 0.00 |
| | 80 | 6.32E-09 | 4.01 | 1.82E-08 | 4.01 | 6.32E-09 | 4.01 | 1.82E-08 | 4.01 | 6.25 |
| | 160 | 3.94E-10 | 4.00 | 1.14E-09 | 3.99 | 3.94E-10 | 4.00 | 1.14E-09 | 3.99 | 4.38 |

TABLE 4.1. Accuracy table for the linear diffusion equation in Sec. 4.1.1, $\tau = 10h$.

4.1.2. *Linear convection diffusion equation.* Consider the linear equation with $\Phi(u) = u, \kappa = 1$ with periodic boundary condition and a source term $s(x)$ added to the right-hand side of the equation (1.1), where

$$s(x) = 8\pi \cos(8\pi x) + (8\pi)^2 \sin(8\pi x).$$

The equation has an exact time-independent solution $u(x, t) = \sin(8\pi x)$. We take $m$ and $M$ in the limiter as $-1$ and 1. The ending time is $T_{end} = 0.2$. The accuracy of the numerical results are presented in Table 4.2. We again observe that the limiter does not affect the designed order of accuracy.

| | without the limiter | | | | | with the limiter | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| $P^1$ | mesh | $L^1$ error | order | $L^\infty$ error | order | $L^1$ error | order | $L^\infty$ error | order | MPP (%) |
| | 10 | 2.78E-01 | | 8.49E-01 | | 2.75E-01 | | 7.05E-01 | | 60.00 |
| | 20 | 8.05E-02 | 1.79 | 2.47E-01 | 1.78 | 8.04E-02 | 1.78 | 2.14E-01 | 1.72 | 40.00 |
| | 40 | 2.09E-02 | 1.95 | 6.43E-02 | 1.94 | 2.09E-02 | 1.95 | 4.89E-02 | 2.13 | 40.00 |
| | 80 | 5.23E-03 | 2.00 | 1.63E-02 | 1.98 | 5.22E-03 | 2.00 | 1.63E-02 | 1.59 | 18.57 |
| | 160 | 1.31E-03 | 2.00 | 4.10E-03 | 1.99 | 1.31E-03 | 2.00 | 4.10E-03 | 1.99 | 8.57 |
| $P^2$ | mesh | $L^1$ error | order | $L^\infty$ error | order | $L^1$ error | order | $L^\infty$ error | order | MPP (%) |
| | 10 | 4.86E-02 | | 2.25E-01 | | 5.02E-02 | | 2.17E-01 | | 40.00 |
| | 20 | 6.86E-03 | 2.83 | 3.18E-02 | 2.82 | 6.86E-03 | 2.87 | 3.18E-02 | 2.77 | 0.00 |
| | 40 | 8.51E-04 | 3.01 | 3.97E-03 | 3.00 | 8.51E-04 | 3.01 | 3.97E-03 | 3.00 | 0.00 |
| | 80 | 1.10E-04 | 2.95 | 5.12E-04 | 2.96 | 1.12E-04 | 2.93 | 5.12E-04 | 2.96 | 10.00 |
| | 160 | 1.37E-05 | 3.00 | 6.44E-05 | 2.99 | 1.38E-05 | 3.02 | 6.44E-05 | 2.99 | 5.00 |
| $P^3$ | mesh | $L^1$ error | order | $L^\infty$ error | order | $L^1$ error | order | $L^\infty$ error | order | MPP (%) |
| | 10 | 7.19E-03 | | 4.21E-02 | | 7.19E-03 | | 4.21E-02 | | 0.00 |
| | 20 | 4.84E-04 | 3.89 | 2.81E-03 | 3.91 | 4.58E-04 | 3.97 | 2.99E-03 | 3.82 | 20.00 |
| | 40 | 3.20E-05 | 3.92 | 1.84E-04 | 3.93 | 3.20E-05 | 3.84 | 1.84E-04 | 4.02 | 13.33 |
| | 80 | 1.99E-06 | 4.01 | 1.15E-05 | 4.00 | 1.99E-06 | 4.01 | 1.15E-05 | 4.00 | 11.25 |
| | 160 | 1.24E-07 | 4.00 | 7.22E-07 | 3.99 | 1.24E-07 | 4.00 | 7.22E-07 | 3.99 | 12.19 |

TABLE 4.2. Accuracy table for the linear convection diffusion equation in Sec. 4.1.2, $\tau = 5h$.

4.1.3. *Nonlinear diffusion equation.* Consider $\Phi(u) \equiv 0, \kappa(u) = (u+1)^2$ in (1.1) with periodic boundary condition and a source term added to the right-hand side of the equation

$$s(x,t) = (8\pi)^2\Big(\sin^3(8\pi x) + 4\sin^2(8\pi x) + 4\sin(8\pi x)\Big) - 2\Big(8\pi\cos(8\pi x)\Big)^2\Big(\sin(8\pi x) + 2\Big).$$

The equation has the steady exact solution $u(x,t) = \sin(8\pi x) + 1$. We compute the solution at $T_{end} = 0.1$ and the numerical results are shown in Table 4.3. For this problem, if we do not use the limiter, then even if we take $\tau/h^2$ as large as $10^6$, the numerical solution still becomes negative. However, with the limiter the scheme is MPP with the designed high order accuracy.

4.1.4. *Nonlinear convection-diffusion equation.* Consider the equation with $\Phi(u) = u^2/2, \kappa(u) = 2(u+1)$ in (1.1). The periodic boundary condition is utilized. And a source term $s(x)$ is added to the right-hand side of the equation, where

$$s(x) = (8\pi)^2\Big(\sin^3(8\pi x) + 4\sin^2(8\pi x) + 4\sin(8\pi x)\Big) - 2\Big(8\pi\cos(8\pi x)\Big)^2\Big(\sin(8\pi x) + 2\Big).$$

Then the equation has an exact steady-state solution $u(x,t) = \sin(8\pi x) + 1$. The global minimum $m$ and maximum $M$ are taken as 0 and 2. All of the results are computed by taking $\tau = 10\kappa_{\max}h$ where $\kappa_{\max} = \max_u(u, 2(u+1)) = 4$. The ending time is taken as $T_{end} = 0.1$ and the results are shown in Table 4.4. We can see that the MPP property is achieved by the limiter and the designed high order accuracy is not affected.

| | without the limiter | | | | with the limiter | | | | |
|---|---|---|---|---|---|---|---|---|---|
| $P^1$ | mesh | $L^1$ error | order | $L^\infty$ error | order | $L^1$ error | order | $L^\infty$ error | order | MPP (%) |
| | 10 | 6.40E-01 | | 1.08E-00 | | 6.06E-01 | | 1.00E-00 | | 80.00 |
| | 20 | 1.71E-01 | 1.90 | 2.46E-01 | 2.14 | 1.69E-01 | 1.84 | 1.99E-01 | 2.33 | 40.00 |
| | 40 | 4.26E-02 | 2.01 | 6.75E-02 | 1.86 | 4.24E-02 | 1.99 | 5.70E-02 | 1.80 | 40.00 |
| | 80 | 1.05E-02 | 2.02 | 1.66E-02 | 2.02 | 1.05E-02 | 2.01 | 1.66E-02 | 1.78 | 19.00 |
| | 160 | 2.62E-03 | 2.01 | 4.12E-03 | 2.01 | 2.62E-03 | 2.01 | 4.12E-03 | 2.01 | 9.00 |
| $P^2$ | mesh | $L^1$ error | order | $L^\infty$ error | order | $L^1$ error | order | $L^\infty$ error | order | MPP (%) |
| | 10 | 1.16E-01 | | 2.91E-01 | | 1.15E-01 | | 2.65E-01 | | 40.00 |
| | 20 | 1.57E-02 | 2.89 | 3.28E-02 | 3.15 | 1.57E-02 | 2.87 | 3.28E-02 | 3.01 | 0.00 |
| | 40 | 1.77E-03 | 3.15 | 4.38E-03 | 2.91 | 1.77E-03 | 3.15 | 4.38E-03 | 2.91 | 10.00 |
| | 80 | 2.21E-04 | 3.01 | 5.37E-04 | 3.03 | 2.26E-04 | 2.97 | 5.37E-04 | 3.03 | 13.75 |
| | 160 | 2.75E-05 | 2.91 | 6.60E-05 | 3.03 | 2.76E-05 | 3.03 | 6.60E-05 | 3.03 | 6.67 |
| $P^3$ | mesh | $L^1$ error | order | $L^\infty$ error | order | $L^1$ error | order | $L^\infty$ error | order | MPP (%) |
| | 10 | 1.78E-02 | | 6.13E-02 | | 1.78E-02 | | 6.13E-02 | | 0.00 |
| | 20 | 1.13E-03 | 3.98 | 2.58E-03 | 4.57 | 1.01E-03 | 4.02 | 2.58E-03 | 4.57 | 20.00 |
| | 40 | 6.61E-05 | 4.09 | 1.88E-04 | 3.78 | 6.61E-05 | 4.06 | 1.88E-04 | 3.78 | 0.00 |
| | 80 | 4.02E-06 | 4.04 | 1.20E-05 | 3.98 | 4.02E-06 | 4.04 | 1.20E-05 | 3.98 | 5.00 |
| | 160 | 2.49E-07 | 4.01 | 7.31E-07 | 4.03 | 2.49E-07 | 4.01 | 7.31E-07 | 4.03 | 3.75 |

TABLE 4.3. Accuracy table for the nonlinear diffusion equation in Sec. 4.1.3, $\tau = 10h$.

| | without the limiter | | | | with the limiter | | | | |
|---|---|---|---|---|---|---|---|---|---|
| $P^1$ | mesh | $L^1$ error | order | $L^\infty$ error | order | $L^1$ error | order | $L^\infty$ error | order | MPP (%) |
| | 10 | 3.17E-01 | | 1.07E+00 | | 3.00E-01 | | 1.00E+00 | | 80.00 |
| | 20 | 8.50E-02 | 1.90 | 2.43E-01 | 2.14 | 8.38E-02 | 1.84 | 1.98E-01 | 2.34 | 40.00 |
| | 40 | 2.12E-02 | 2.00 | 6.71E-02 | 1.85 | 2.12E-02 | 1.99 | 5.68E-02 | 1.80 | 40.00 |
| | 80 | 5.26E-03 | 2.01 | 1.66E-02 | 2.02 | 5.26E-03 | 2.01 | 1.66E-02 | 1.78 | 20.00 |
| | 160 | 1.31E-03 | 2.00 | 4.12E-03 | 2.01 | 1.31E-03 | 2.00 | 4.12E-03 | 2.01 | 8.75 |
| $P^2$ | mesh | $L^1$ error | order | $L^\infty$ error | order | $L^1$ error | order | $L^\infty$ error | order | MPP (%) |
| | 10 | 5.79E-02 | | 2.90E-01 | | 5.72E-02 | | 2.61E-01 | | 40.00 |
| | 20 | 7.83E-03 | 2.89 | 3.27E-02 | 3.15 | 7.83E-03 | 2.87 | 3.27E-02 | 3.00 | 0.00 |
| | 40 | 8.85E-04 | 3.14 | 4.37E-03 | 2.90 | 8.82E-04 | 3.15 | 4.37E-03 | 2.90 | 10.00 |
| | 80 | 1.10E-04 | 3.01 | 5.37E-04 | 3.03 | 1.13E-04 | 2.97 | 5.37E-04 | 3.03 | 15.00 |
| | 160 | 1.37E-05 | 3.01 | 6.60E-05 | 3.02 | 1.38E-05 | 3.03 | 6.60E-05 | 3.02 | 6.67 |
| $P^3$ | mesh | $L^1$ error | order | $L^\infty$ error | order | $L^1$ error | order | $L^\infty$ error | order | MPP (%) |
| | 10 | 8.89E-03 | | 6.09E-02 | | 8.89E-03 | | 6.09E-02 | | 0.00 |
| | 20 | 5.62E-04 | 3.98 | 2.58E-03 | 4.56 | 5.48E-04 | 4.02 | 2.58E-03 | 4.56 | 20.00 |
| | 40 | 3.30E-05 | 4.09 | 1.88E-04 | 3.78 | 3.30E-05 | 4.06 | 1.88E-04 | 3.78 | 0.00 |
| | 80 | 2.01E-06 | 4.04 | 1.19E-05 | 3.97 | 2.01E-06 | 4.04 | 1.19E-05 | 3.97 | 5.00 |
| | 160 | 1.25E-07 | 4.01 | 7.30E-07 | 4.03 | 1.25E-07 | 4.01 | 7.30E-07 | 4.03 | 3.75 |

TABLE 4.4. Accuracy table for the nonlinear convection-diffusion equation in Sec. 4.1.4, $\tau = 10\kappa_{max}h$.

4.2. **1D porous medium equation.** Consider the porous medium equation which means $\Phi = 0, \kappa(u) = nu^{n-1}$ in (1.1) with periodic boundary condition. Here we take $\Omega = [-6, 6]$. $n$ is a positive integer and we will mainly consider the cases of $n = 2, 3, 5, 8$. For this equation, we have a famous Barenblatt solution which is defined as follows:

$$B_n(x,t) = t^{-z}\left(\left(1 - \frac{z(n-1)}{2n}\frac{|x|^2}{t^{2z}}\right)_+\right)^{\frac{1}{n-1}},$$

where $z = (n+1)^{-1}$ and $(a)_+ = \max(a, 0)$. We take $B_n(x, 1)$ as the initial condition and compute the solution at $T_{end} = 2$. This equation is much more difficult in numerical computation compared with the previous examples because of its degeneracy, which means that $\kappa(u)$ may be equal to zero somewhere. When $\kappa(u) = 0$, the property of the equation will change and the solution will be more complicated. Furthermore, in this equation, the exact solution grows rapidly but not smoothly from zero to one which could generate numerical oscillation and negative value easily. However the negative solution will lose its physical meaning and will make the equation ill-posed for odd $n$. Therefore, the maximum-principle-preserving, especially the positive-preserving property is important. The numerical results computed by the LDG scheme with the scaling MPP limiter on 160 cells are presented in Figure 4.1 and the minimum of the numerical solution is exactly zero. The time step $\tau$ is taken from $5h^2$ to $10h^2$ where $h$ is the spatial step. The comparison of results of the implicit LDG scheme with and without the MPP limiter is shown in Figure 4.2.

4.3. **The 1D Buckley-Leverett equation.** Consider the Buckley-Leverett equation which means that

$$\Phi(u) = \frac{u^2}{u^2 + (1-u)^2}, \quad \nu(u) = 4\epsilon u(1-u),$$

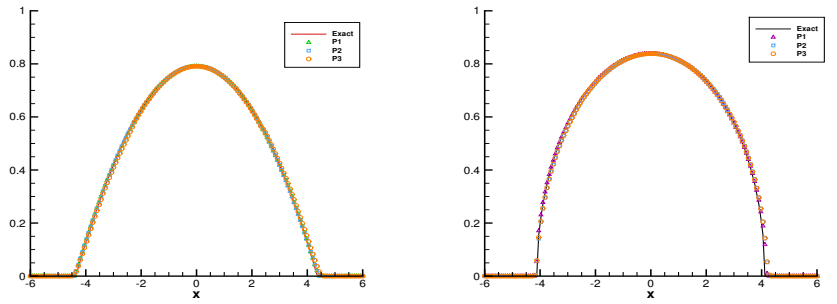
where $\epsilon = 0.01$ in (1.1) and $\Omega = [0, 1]$. The Dirichlet boundary condition $u(0, t) = 1$ is used. The initial condition is

$$u(x,0) = \begin{cases} 1 - 3x, & x \in \left(0, \frac{1}{3}\right), \\ 0, & else. \end{cases}$$

The final time $T_{end} = 0.2$. The time step $\tau$ is taken from $0.1h^2$ to $2h^2$. The solution solved by an explicit MPP LDG scheme on 400 cells is used as the reference. The numerical solutions computed by $P^1$, $P^2$ and $P^3$ LDG on 100 cells are shown in Figure 4.3 and the percentages of usage of the MPP limiter are 2.74%, 6.45%, 2.74% respectively in the $P^1$, $P^2$ and $P^3$ cases.

4.4. **2D accuracy test.** In this section, we consider the accuracy of the implicit LDG scheme with the scaling MPP limiter. We again take $\Omega = [0, 1] \times [0, 1]$.

4.4.1. *Linear equation.* Firstly, we validate the accuracy for solving the linear diffusion equation, i.e., $\Phi = 0$ and $\kappa = 1$ in (1.1) with periodic boundary condition and a source term $s(x, y) = 128\pi^2 \sin(8\pi(x + y))$ on the right hand of the equation. This problem has the exact steady solution $u = \sin(8\pi(x + y))$. The ending time is taken as $T_{end} = 0.1$. The numerical results are shown in Table 4.5. We can clearly see that the limiter does




(a) n = 2 (b) n = 3

(c) n = 5 (d) n = 8

FIGURE 4.1. The numerical solution on 160 cells of the 1D porous medium equation in Sec. 4.2.

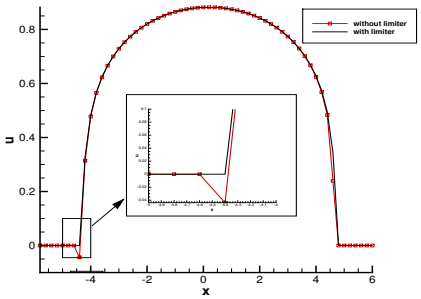



FIGURE 4.2. Comparison between the implicit $P^1$ scheme with and without the MPP limiter on 60 cells for the porous medium equation where $n = 5$ in Sec. 4.2.

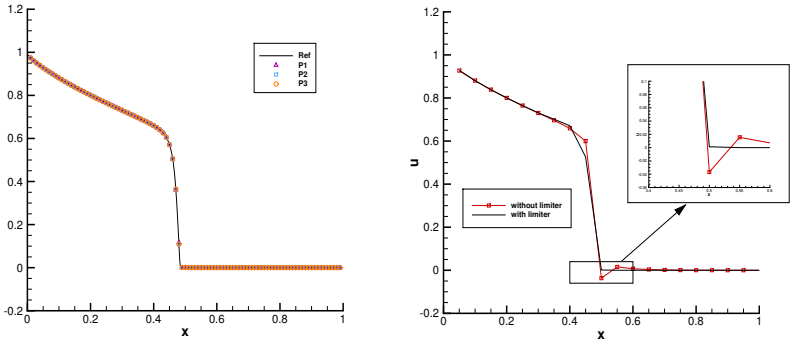



FIGURE 4.3. The numerical solution of the 1D Buckley-Leverett equation in Sec. 4.3. Left: the MPP $P^1$, $P^2$ and $P^3$ schemes on 100 cells; Right: comparison of the $P^1$ LDG scheme with and without the MPP limiter on 20 cells.

not affect the designed order of accuracy. Notice that, even though we prove the MPP property only for the second order $Q^1$ scheme, it seems that the MPP property holds for higher order cases as well.

| | without the limiter | | | | | with the limiter | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| $P^1$ | mesh | $L^1$ error | order | $L^\infty$ error | order | $L^1$ error | order | $L^\infty$ error | order | MPP (%) |
| | 10 | 3.54E-01 | | 1.87E-00 | | 2.63E-01 | | 1.00E-00 | | 100.00 |
| | 20 | 1.44E-01 | 1.30 | 5.35E-01 | 1.80 | 1.40E-01 | 0.91 | 5.87E-01 | 0.77 | 80.00 |
| | 40 | 4.03E-02 | 1.84 | 1.33E-01 | 2.01 | 4.03E-02 | 1.80 | 1.72E-01 | 1.78 | 60.00 |
| | 80 | 1.04E-02 | 1.96 | 3.28E-02 | 2.01 | 1.04E-02 | 1.96 | 4.94E-02 | 1.80 | 27.50 |
| | 160 | 2.61E-03 | 1.99 | 8.22E-03 | 2.00 | 2.61E-03 | 1.99 | 1.26E-02 | 1.97 | 10.00 |
| $P^2$ | mesh | $L^1$ error | order | $L^\infty$ error | order | $L^1$ error | order | $L^\infty$ error | order | MPP (%) |
| | 10 | 6.41E-02 | | 4.64E-01 | | 5.01E-02 | | 3.91E-01 | | 100.00 |
| | 20 | 8.95E-03 | 2.84 | 6.24E-02 | 2.90 | 8.95E-03 | 2.48 | 6.24E-02 | 2.65 | 0.00 |
| | 40 | 1.12E-03 | 2.99 | 7.90E-03 | 2.98 | 1.14E-03 | 2.98 | 7.90E-03 | 2.98 | 20.00 |
| | 80 | 1.41E-04 | 3.00 | 1.02E-03 | 2.95 | 1.41E-04 | 3.01 | 1.02E-03 | 2.95 | 20.00 |
| | 160 | 1.75E-05 | 3.00 | 1.29E-04 | 2.99 | 1.78E-05 | 3.01 | 1.29E-04 | 2.99 | 10.00 |
| $P^3$ | mesh | $L^1$ error | order | $L^\infty$ error | order | $L^1$ error | order | $L^\infty$ error | order | MPP (%) |
| | 10 | 9.39E-03 | | 8.20E-02 | | 9.66E-03 | | 8.20E-02 | | 20.00 |
| | 20 | 6.25E-04 | 3.91 | 5.60E-03 | 3.87 | 6.11E-04 | 3.98 | 5.60E-03 | 3.87 | 20.00 |
| | 40 | 3.91E-05 | 4.00 | 3.68E-04 | 3.93 | 3.91E-05 | 3.97 | 3.68E-04 | 3.93 | 0.00 |
| | 80 | 2.44E-06 | 4.00 | 2.29E-05 | 4.00 | 2.38E-06 | 4.04 | 2.29E-05 | 4.00 | 15.00 |
| | 160 | 1.52E-07 | 4.00 | 1.45E-06 | 3.99 | 1.50E-07 | 3.98 | 1.45E-06 | 3.99 | 7.50 |

TABLE 4.5. Accuracy table for the 2D diffusion equation in Sec. 4.4.1, $\tau = 10h$.

4.4.2. *Nonlinear equation.* Consider the nonlinear diffusion equation with $\Phi = 0$ and $\kappa(u) = 2(u+1)$ in (1.1). The periodic boundary condition is used and a source term

$s(x,y)$ is added to the right-hand side of (1.1), where

$$s(x,y) = 16\pi[2\sin(8\pi(x+y)) + 8\sin(8\pi(x+y))^2 + 3\sin(8\pi(x+y))^3 - 4].$$

The exact steady-state solution is $u(x,y) = \frac{1}{8\pi}(1+\sin(8\pi(x+y)))$ and the results are presented in Table 4.6. Again, we observe our scheme is MPP and the limiter does not affect the designed order of accuracy.

| | without the limiter | | | | | with the limiter | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| $P^1$ | mesh | $L^1$ error | order | $L^\infty$ error | order | $L^1$ error | order | $L^\infty$ error | order | MPP (%) |
| | 10 | 1.40E-02 | | 7.36E-02 | | 1.04E-02 | | 3.98E-02 | | 100.00 |
| | 20 | 5.73E-03 | 1.29 | 2.14E-02 | 1.78 | 5.58E-03 | 0.91 | 2.34E-02 | 0.77 | 80.00 |
| | 40 | 1.60E-03 | 1.84 | 5.28E-03 | 2.02 | 1.60E-03 | 1.80 | 6.87E-03 | 1.78 | 60.00 |
| | 80 | 4.13E-04 | 1.96 | 1.31E-03 | 2.02 | 4.13E-04 | 1.96 | 1.97E-03 | 1.81 | 27.50 |
| | 160 | 1.04E-04 | 1.99 | 3.27E-04 | 2.00 | 1.04E-04 | 1.99 | 5.02E-04 | 1.97 | 13.75 |
| $P^2$ | mesh | $L^1$ error | order | $L^\infty$ error | order | $L^1$ error | order | $L^\infty$ error | order | MPP (%) |
| | 10 | 2.55E-03 | | 1.86E-2 | | 2.01E-03 | | 1.55E-02 | | 100.00 |
| | 20 | 3.56E-04 | 2.84 | 2.47E-03 | 2.91 | 3.56E-04 | 2.50 | 2.47E-03 | 2.65 | 0.00 |
| | 40 | 4.47E-05 | 2.99 | 3.16E-04 | 2.96 | 4.53E-05 | 2.97 | 3.16E-04 | 2.96 | 20.00 |
| | 80 | 5.59E-06 | 3.00 | 4.08E-05 | 2.95 | 5.62E-06 | 3.01 | 4.08E-05 | 2.95 | 23.33 |
| | 160 | 6.97E-07 | 3.01 | 5.13E-06 | 2.99 | 6.99E-07 | 3.01 | 5.13E-06 | 2.99 | 11.25 |
| $P^3$ | mesh | $L^1$ error | order | $L^\infty$ error | order | $L^1$ error | order | $L^\infty$ error | order | MPP (%) |
| | 10 | 3.74E-04 | | 3.26E-03 | | 3.85E-04 | | 3.26E-03 | | 20.00 |
| | 20 | 2.49E-05 | 3.91 | 2.21E-04 | 3.88 | 2.43E-05 | 3.98 | 2.21E-04 | 3.88 | 20.00 |
| | 40 | 1.55E-06 | 4.00 | 1.47E-05 | 3.91 | 1.55E-06 | 3.97 | 1.47E-05 | 3.91 | 0.00 |
| | 80 | 9.71E-08 | 4.00 | 9.14E-07 | 4.00 | 9.48E-08 | 4.03 | 9.48E-07 | 3.95 | 10.00 |
| | 160 | 6.06E-09 | 4.00 | 5.75E-08 | 3.99 | 5.98E-09 | 3.99 | 5.90E-08 | 4.00 | 6.25 |

TABLE 4.6. Accuracy table for the 2D nonlinear equation in Sec. 4.4.2, $\tau = 10h$.

4.5. **The 2D porous medium equation.** Consider the following equation with periodic boundary condition

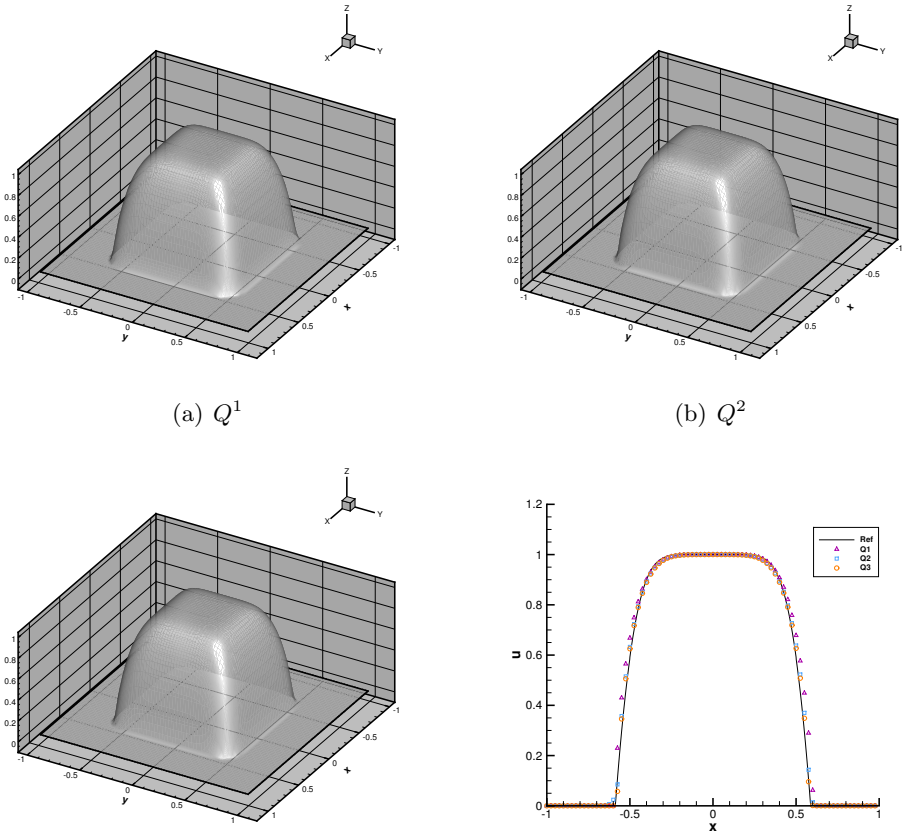
$$u_t = \nabla \cdot (\kappa(u)\nabla u),$$

where $\kappa(u) = 2u$ and the initial condition is

$$u(x,y,0) = \begin{cases} 0, & (x,y) \in \left[-\frac{1}{2},\frac{1}{2}\right] \times \left[-\frac{1}{2},\frac{1}{2}\right], \\ 1, & else. \end{cases} \tag{4.1}$$

The result at $T = 0.005$ is shown in Figure 4.4 and the MPP limiter works well. The time step $\tau$ is taken from $0.001h^2$ to $0.02h^2$. The reference solution is computed by an explicit method (RK2) in time and central difference method in [17] in space on a $1000 \times 1000$ mesh.

Figure 4.5 is the cut at $x = 0$ of the $P^1$ LDG numerical solution to compare the results with and without the MPP limiter on a $20 \times 20$ mesh.




(a) $Q^1$ (b) $Q^2$

(c) $Q^3$ (d) cut along $x = 0$

FIGURE 4.4. The numerical solution of the $P^1$, $P^2$ and $P^3$ LDG method for the 2D porous medium equation in Sec. 4.5 at $T = 0.005$ on $80 \times 80$ cells.

## 5. Concluding remarks

In this paper, we present two types of theoretical results about the MPP property of implicit LDG schemes for solving linear parabolic or convection-diffusion equations with periodic boundary conditions. The first type enjoys high order accuracy, but is restricted to uniform meshes and pure diffusion equations. The other type is only for second order accuracy in 1D, but can handle nonuniform meshes and convection-diffusion equations. These two types of results are summarized as the following Table 5. Numerical experiments suggest that the scaling MPP limiter works well to achieve the MPP property for nonlinear equations as well. There is still much left to be done in terms of the MPP property for implicit DG schemes. In the future, we would like to pursue further research in the following three directions. Firstly, high order time discretizations instead of backward Euler scheme will be considered. Secondly, we hope to extend the work to triangular grids and expect that the scaling MPP limiter could loosen the restrictions on

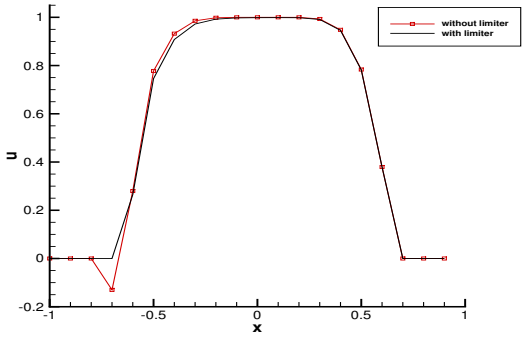



FIGURE 4.5. The numerical solution of the 2D porous medium equation in Sec. 4.5 computed by the implicit LDG with and without the limiter. The mesh size $h = 0.05$.

| | linear diffusion equation (Sec. 2) | linear convection diffusion equation (Sec. 3) |
|---|---|---|
| | lower bounds $\sigma_{\min}$ with the limiter | lower bounds $\sigma_{\min}$ without the limiter |
| 1D $P^1$ | 0.0556 | 15.190 |
| $P^2$ | 0.1620 | - |
| $P^3$ | 0.0544 | - |
| $P^5$ | 0.0379 | - |
| $P^7$ | 0.0238 | - |
| $P^9$ | 0.0157 | - |
| $\vdots$ | $\vdots$ | - |
| 2D $Q^1$ | 1.0000 | - |

TABLE 5.1. Overview of the main content of this article.

the mesh. Finally, we would pursue the introduction of new methods to analyze quasi-linear parabolic equations and more general boundary conditions, such as the Dirichlet and Neumann boundary conditions.

## REFERENCES

1. Santiago Badia, Jesús Bonilla, and Alba Hierro, *Differentiable monotonicity-preserving schemes for discontinuous Galerkin methods on arbitrary meshes*, Computer Methods in Applied Mechanics and Engineering **320** (2017), 582–605.
2. Santiago Badia and Alba Hierro, *On discrete maximum principles for discontinuous Galerkin methods*, Computer Methods in Applied Mechanics and Engineering **286** (2015), 107–122.
3. Gabriel R. Barrenechea, Volker John, and Petr Knobloch, *Finite element methods respecting the discrete maximum principle for convection-diffusion equations*, SIAM Review **66** (2024), 3–88.
4. Abraham Berman and Robert J. Plemmons, *Chapter 6 - M-matrices*, Nonnegative Matrices in the Mathematical Sciences (Abraham Berman and Robert J. Plemmons, eds.), Academic Press, 1979, pp. 132–164.
5. François Bouchon, *Monotonicity of some perturbations of irreducibly diagonally dominant M-matrices*, Numerische Mathematik **105** (2007), 591–601.

6. A. Carmona, A.M. Encinas, S. Gago, M.J. Jiménez, and M. Mitjana, *The inverses of some circulant matrices*, Applied Mathematics and Computation **270** (2015), 785–793.
7. Juan Cheng and Chi-Wang Shu, *Positivity-preserving Lagrangian scheme for multi-material compressible flow*, Journal of Computational Physics **257** (2014), 143–168.
8. Bernardo Cockburn, Suchung Hou, and Chi-Wang Shu, *The Runge-Kutta local projection discontinuous Galerkin finite element method for conservation laws. IV. the multidimensional case*, Mathematics of Computation **54** (1990), 545–581.
9. Bernardo Cockburn, George E Karniadakis, and Chi-Wang Shu, *The development of discontinuous Galerkin methods*, Discontinuous Galerkin Methods: Theory, Computation and Applications, Springer, 2000, pp. 3–50.
10. Bernardo Cockburn, San-Yih Lin, and Chi-Wang Shu, *TVB Runge-Kutta local projection discontinuous Galerkin finite element method for conservation laws III: One-dimensional systems*, Journal of Computational Physics **84** (1989), 90–113.
11. Bernardo Cockburn and Chi-Wang Shu, *TVB Runge-Kutta local projection discontinuous Galerkin finite element method for conservation laws. II: General framework*, Mathematics of Computation **52** (1989), 411–435.
12. ———, *The local discontinuous Galerkin method for time-dependent convection-diffusion systems*, SIAM Journal on Numerical Analysis **35** (1998), 2440–2463.
13. ———, *The Runge-Kutta discontinuous Galerkin method for conservation laws V: Multidimensional systems*, Journal of Computational Physics **141** (1998), 199–224.
14. ———, *The Runge-Kutta local projection $P^1$ discontinuous Galerkin finite element method for scalar conservation laws*, ESAIM: M2AN **25** (1991), 337–361.
15. ———, *Runge–Kutta discontinuous Galerkin methods for convection-dominated problems*, Journal of Scientific Computing **16** (2001), 173–261.
16. Jie Du and Yang Yang, *Maximum-principle-preserving third-order local discontinuous Galerkin method for convection-diffusion equations on overlapping meshes*, Journal of Computational Physics **377** (2019), 117–141.
17. Steinar Evje and Kenneth Hvistendahl Karlsen, *Monotone difference approximations of BV solutions to degenerate convection-diffusion equations*, SIAM Journal on Numerical Analysis **37** (2000), 1838–1860.
18. Sigal Gottlieb and Chi-Wang Shu, *Total variation diminishing Runge-Kutta schemes*, Mathematics of Computation **67** (1998), 73–85.
19. Sigal Gottlieb, Chi-Wang Shu, and Eitan Tadmor, *Strong stability-preserving high-order time discretization methods*, SIAM Review **43** (2001), 89–112.
20. Li Guo and Yang Yang, *Positivity preserving high-order local discontinuous Galerkin method for parabolic equations with blow-up solutions*, Journal of Computational Physics **289** (2015), 181–195.
21. Tamas Horvath and M. Mincsovics, *Discrete maximum principle for interior penalty discontinuous Galerkin methods*, Central European Journal of Mathematics **11** (2013), 664–679.
22. Charles R. Johnson, Ronald L. Smith, and Michael J. Tsatsomeros, *Matrix positivity*, Cambridge University Press, 2020, pp.90–163.
23. Hao Li and Xiangxiong Zhang, *On the monotonicity and discrete maximum principle of the finite difference implementation of $C^0$-$Q^2$ finite element method*, Numerische Mathematik **145** (2020), 437–472.
24. Tong Qin and Chi-Wang Shu, *Implicit positivity-preserving high-order discontinuous Galerkin methods for conservation laws*, SIAM Journal on Scientific Computing **40** (2018), A81–A107.
25. Jie Shen, Tao Tang, and Li-Lian Wang, *Spectral methods: algorithms, analysis and applications*, Springer, 2011, pp.108–113.
26. Vidar Thomée and Lars Wahlbin, *On the existence of maximum principles in parabolic finite element equations*, Mathematics of Computation **77** (2008), 11–19.
27. J. J. W. van der Vegt, Yinhua Xia, and Yan Xu, *Positivity preserving limiters for time-implicit higher order accurate discontinuous Galerkin discretizations*, SIAM Journal on Scientific Computing **41** (2019), A2037–A2063.

28. Tomás Vejchodský and Pavel Solín, *Discrete maximum principle for higher-order finite elements in 1D*, Mathematics of Computation **76** (2007), 1833–1846.
29. Tomáš Vejchodský, *Higher-order discrete maximum principle for 1D diffusion–reaction problems*, Applied Numerical Mathematics **60** (2010), 486–500.
30. Tao Xiong, Jing-Mei Qiu, and Zhengfu Xu, *High order maximum-principle-preserving discontinuous Galerkin method for convection-diffusion equations*, SIAM Journal on Scientific Computing **37** (2015), A583–A608.
31. Jinchao Xu and Ludmil T. Zikatanov, *A monotone finite element scheme for convection-diffusion equations*, Mathematics of Computation **68** (1999), 1429–1446.
32. Zhengfu Xu and Xiangxiong Zhang, *Bound-preserving high-order schemes*, Handbook of Numerical Analysis **18** (2017), 81–102.
33. Ziyao Xu and Chi-Wang Shu, *Third order maximum-principle-satisfying and positivity-preserving Lax-Wendroff discontinuous Galerkin methods for hyperbolic conservation laws*, Journal of Computational Physics **470** (2022), 111591.
34. Qiang Zhang, Zi-long Wu, *Numerical simulation for porous medium equation by local discontinuous Galerkin finite element method*, Journal of Scientific Computing **38** (2009), 127–148.
35. Xiangxiong Zhang and Chi-Wang Shu, *On maximum-principle-satisfying high order schemes for scalar conservation laws*, Journal of Computational Physics **229** (2010), 3091–3120.
36. ———, *On positivity-preserving high order discontinuous Galerkin schemes for compressible euler equations on rectangular meshes*, Journal of Computational Physics **229** (2010), 8918–8934.
37. Xiangxiong Zhang, Yinhua Xia, and Chi-Wang Shu, *Maximum-principle-satisfying and positivity-preserving high order discontinuous Galerkin schemes for conservation laws on triangular meshes*, Journal of Scientific Computing **50** (2011), 29–62.
38. Yifan Zhang, Xiangxiong Zhang, and Chi-Wang Shu, *Maximum-principle-satisfying second order discontinuous Galerkin schemes for convection-diffusion equations on triangular meshes*, Journal of Computational Physics **234** (2013), 295–316.

## Appendix A. Legendre polynomials and $\delta_y^k$ polynomials

We consider the standard Legendre polynomials $\{p_n(x)\}_{n=0}^{\infty}$ on $\hat{I}=[-1,1]$ generated by the following recursive relation:

$$(n+1)p_{n+1}(x)=(2n+1)xp_n(x)-np_{n-1}(x),\quad p_0(x)=1,\quad p_1(x)=x,\quad \forall x\in\hat{I}.$$

They have many properties (see, e.g. [25]) and we only present a few here in the following lemma.

**Lemma A.1.** *Denote $p^{(i)}(x)$ as the i-th derivative of $p(x)$ and $n!$ as the factorial of $n$. Legendre polynomials enjoy the following properties which will be useful for our proof.*
*(1)* $p_n(1)=1,\ p_n^{(i)}(-x)=(-1)^{n+i}p_n^{(i)}(x)\ \forall x\in\hat{I}$, *and* $|p_n(x)|<1,\ \forall x\in\hat{I}$.
*(2)* $(2n+1)p_n(x)=p'_{n+1}(x)-p'_{n-1}(x)$.
*(3)* $p_n^{(i)}(1)=\frac{2^{-i}}{i!}\frac{(n+i)!}{(n-i)!}$, *where* $i=0,\cdots,n$.
*(4)* $p_n^{(i)}(1)-p_n^{(i)}(x)>0$, *where* $i=0,\cdots,n-1$.
*(5)* $\int_{\hat{I}}p_{n_1}(x)p_{n_2}(x)dx=\frac{2}{2n_1+1}\delta_{n_1n_2}$, *where* $\delta_{n_1n_2}$ *is Kronecker delta.*
*(6)* $p_n^{(n)}(1)\geq p_n^{(i)}(1)(n-i)!$, *where* $i=0,1,\cdots,n-1$.
*(7)* $p_l^{(i)}(1)>p_{l-n}^{(i)}(1)$, *where* $n=1,2,\cdots,l$ *and* $i=0,\cdots,l$.
*(8)* $p_n^{(i)}(1)\geq|p_n^{(i)}(x)|$, *where* $i=0,\cdots,n$.

The following lemma shows that $\delta_y^k(x)$ has similar property to the Dirac delta distribution in $P^k(\hat{I})$.

**Lemma A.2.** *$\delta_y^k(x)$ has the following properties:*
*(1) For any fixed $y \in \hat{I}$, $\delta_y^k(x) \in P^k(\hat{I})$ and $(w(x), \delta_y^k(x))_{\hat{I}} = w(y)$ holds for any $w(x) \in P^k(\hat{I})$.*
*(2) Fix $y = 1$ and define*

$$\delta_{1,j}^k(x) = \frac{2}{h_j}\delta_1^k(T_j(x)),\ x \in I_j.$$

*Then we have $(w(x), \delta_{1,j}^k(x))_j = w(x_{j+\frac{1}{2}})$ for any $w(x) \in P^k(I_j)$. We define $\delta_{-1,j}^k(x)$ in the same way.*
*(3)$(\delta_1^k)^{(i)}(1) - (\delta_1^k)^{(i)}(x) > 0,\ x \in [-1, 1)$ holds for any $k \geq 1$ and $i = 0, \cdots, k-1$.*
*(4) $(\delta_1^k)^{(k-2i)}(-1) > 0,\ (\delta_1^k)^{(k-2i-1)}(-1) < 0$ for $i = 0, \cdots, \lfloor k/2 \rfloor$.*
*(5) $(\delta_1^k)^{(k-2i)}(-1) + (\delta_1^k)^{(k-2i-1)}(1) > 0$ for $i = 0, \cdots, \lfloor k/2 \rfloor$.*
*(6) $(\delta_{1,j}^k)^{(i)}(x) = (\frac{2}{h_j})^{l+1}(\delta_1^k)^{(i)}(x)$, for $i = 0, \cdots, k$.*

**Proof:** All of the above properties can be found in [24]. Interested readers can refer Lemma 3.8 in [24] for more details. □

Furthermore, we need a few new properties in our proof, which are shown below.

**Lemma A.3.** *$(\delta_1^k)^{(i)}(x)$ have the following two other representations besides the original definition:*

$$(\delta_1^k)^{(i)}(x) = \sum_{l=k}^{k+1} \frac{1}{2} p_l^{(i+1)}(x), \tag{A.1}$$

*and*

$$(\delta_1^k)^{(i)}(x) = \sum_{l=k-1}^{k} \frac{1}{2} p_l^{(i+1)}(x) + \frac{2k+1}{2} p_k^{(i)}(x). \tag{A.2}$$

**Proof:** Using the second property in Lemma (A.1), we have

$$\delta_1^{(i)}(x) = \sum_{l=i}^{k} \frac{2l+1}{2} p_l^{(i)}(x) = \sum_{l=i}^{k} \frac{1}{2}\left(p_{l+1}^{(i+1)}(x) - p_{l-1}^{(i+1)}(x)\right)$$
$$= \sum_{l=i+1}^{k+1} \frac{1}{2} p_l^{(i+1)}(x) - \sum_{l=i-1}^{k-1} \frac{1}{2} p_l^{(i+1)}(x) = \sum_{l=i+1}^{k+1} \frac{1}{2} p_l^{(i+1)}(x) - \sum_{l=i+1}^{k-1} \frac{1}{2} p_l^{(i+1)}(x) = \sum_{l=k}^{k+1} \frac{1}{2} p_l^{(i+1)}(x).$$

Using the above equation, we have

$$\delta_1^{(i)}(x) = \sum_{l=i}^{k} \frac{2l+1}{2} p_l^{(i)}(x) = \sum_{l=i}^{k-1} \frac{2l+1}{2} p_l^{(i)}(x) + \frac{2k+1}{2} p_k^{(i)}(x) = \sum_{l=k-1}^{k} \frac{1}{2} p_l^{(i+1)}(x) + \frac{2k+1}{2} p_k^{(i)}(x).$$

Then we finish proof of this lemma. □

## Appendix B. Proofs in Section 2

B.1. **Proof of Theorem 2.4.** We will split the theorem into the following two lemmas and prove them separately.

**Lemma B.1.** *Denote* $\bar{\mathbf{u}} = (\bar{u}_1, \cdots, \bar{u}_N)^\top$ *and* $\mathbf{f}^{\pm} = (f^{\pm}_{1\mp\frac{1}{2}}, \cdots, f^{\pm}_{N\mp\frac{1}{2}})^\top$*, then we have*

$$\mathbf{D}^{-1}(\mathbf{T}_1\mathbf{T}_3 - \mathbf{T}_2\mathbf{T}_4)\bar{\mathbf{u}} = \frac{\tau}{h}\mathbf{D}^{-1}\Big(2\mathbf{T}_4\mathbf{B}^\top - \mathbf{T}_1\mathbf{A}\Big)\mathbf{f}_1 + \mathbf{T}_1\bar{\mathbf{f}} + \mathbf{T}_4(\mathbf{f}^+ + \mathbf{f}_2), \tag{B.1}$$

*where*

$$\begin{aligned}
\mathbf{f}_1 &= \Big(\sum_{i=0}^{\frac{k-3}{2}} \tau^i (f, \delta_{1,1}^{(2i+1)})_1, \cdots, \sum_{i=0}^{\frac{k-3}{2}} \tau^i (f, \delta_{1,N}^{(2i+1)})_N\Big)^\top, \\
\mathbf{f}_2 &= \Big(\sum_{i=1}^{\frac{k-1}{2}} \tau^i (f, \delta_{-1,1}^{(2i)})_1, \cdots, \sum_{i=1}^{\frac{k-1}{2}} \tau^i (f, \delta_{-1,N}^{(2i)})_N\Big)^\top,
\end{aligned} \tag{B.2}$$

*and the definitions of* $\mathbf{T}_\varrho$ $(\varrho = 1, 2, 3, 4)$, $\mathbf{A}$, $\mathbf{B}$, $\mathbf{D}$, $\xi$, $\eta$, $\xi^*$, $\eta^*$, $\omega^*$, $\gamma^*$ *are shown in Theorem 2.4.*

**Proof:** Take $v = 1$ in (2.4), we have

$$\bar{u}_j = \bar{f}_j + \frac{\tau}{h}\left(q^-_{j+\frac{1}{2}} - q^-_{j-\frac{1}{2}}\right). \tag{B.3}$$

To eliminate $q_{j+\frac{1}{2}}$, we take $w(x) = \delta_{1,j}(x)$ in (2.5) and have

$$q^-_{j+\frac{1}{2}} + (u, \delta'_{1,j})_j - \left(u^+_{j+\frac{1}{2}}\delta_{1,j}(x_{j+\frac{1}{2}}) - u^+_{j-\frac{1}{2}}\delta_{1,j}(x_{j-\frac{1}{2}})\right) = 0. \tag{a.0}$$

Then taking $v = \delta'_{1,j}(x)$ in (2.4), we have

$$(u, \delta'_{1,j})_j + \tau(q, \delta''_{1,j})_j - \tau\left(q^-_{j+\frac{1}{2}}\delta'_{1,j}(x_{j+\frac{1}{2}}) - q^-_{j-\frac{1}{2}}\delta'_{1,j}(x_{j-\frac{1}{2}})\right) = (f, \delta'_{1,j})_j, \tag{a.1}$$

and so on. Because $k$ is odd, we stop at taking $w = \delta^{(k-1)}_{1,j}(x)$ in (2.5):

$$(q, \delta^{(k-1)}_{1,j})_j + (u, \delta^{(k)}_{1,j})_j - \left(u^+_{j+\frac{1}{2}}\delta^{(k-1)}_{1,j}(x_{j+\frac{1}{2}}) - u^+_{j-\frac{1}{2}}\delta^{(k-1)}_{1,j}(x_{j-\frac{1}{2}})\right) = 0. \tag{a.k-1}$$

The procedure can be rewritten in short as below:

$$q^-_{j+\frac{1}{2}} + (u, \delta'_{1,j})_j - K^1_{j,0} = 0, \tag{a.0}$$

$$(u, \delta'_{1,j})_j + \tau(q, \delta''_{1,j})_j - \tau K^2_{j,1} = K^3_{j,1}, \tag{a.1}$$

$$\vdots$$

$$(u, \delta^{(k-2)}_{1,j})_j + \tau(q, \delta^{(k-1)}_{1,j}) - \tau K^2_{j,k-2} = K^3_{j,k-2}, \tag{a.k-2}$$

$$(q, \delta^{(k-1)}_{1,j})_j + (u, \delta^{(k)}_{1,j})_j - K^1_{j,k-1} = 0, \tag{a.k-1}$$

where

$$K^1_{j,l} = u^+_{j+\frac{1}{2}}\delta^{(l)}_{1,j}(x_{j+\frac{1}{2}}) - u^+_{j-\frac{1}{2}}\delta^{(l)}_{1,j}(x_{j-\frac{1}{2}}),\ K^2_{j,l} = q^-_{j+\frac{1}{2}}\delta^{(l)}_{1,j}(x_{j+\frac{1}{2}}) - q^-_{j-\frac{1}{2}}\delta^{(l)}_{1,j}(x_{j-\frac{1}{2}}),$$
$$K^3_{j,l} = (f, \delta^{(l)}_{1,j}(x))_j, \qquad K^4_{j,l,l+1} = K^1_{j,l} - \tau K^2_{j,l+1} - K^3_{j,l+1},$$

and $l$ is an even number less than $k-1$. Then we have the following equation from (a.$l$) and (a.$l$+1)

$$(q, \delta^{(l)}_{1,j})_j = \tau(q, \delta^{(l+2)}_{1,j})_j + K^4_{j,l,l+1}.$$

Then we have

$$q^-_{j+\frac{1}{2}} = -\tau^{\frac{k-1}{2}} h\delta^{(k)}_{1,j}\bar{u}_j + \sum_{l=0}^{\frac{k-3}{2}} \tau^l K^4_{j,2l,2l+1} + \tau^{\frac{k-1}{2}} K^1_{j,k-1}. \tag{B.4}$$

There is a new unknown variable $u^+_{j-\frac{1}{2}}$ which we can get the following expression through the same way. We have

$$u^+_{j-\frac{1}{2}} = -\tau^{\frac{k+1}{2}} h\delta^{(k)}_{-1,j}\bar{q}_j - \left(\tau^{\frac{k-1}{2}} K^{4,*}_{j,k-2,k-1} + \cdots + \tau K^{4,*}_{j,1,2}\right) + \tau K^{2,*}_{j,0} + f^+_{j-\frac{1}{2}} \tag{B.5}$$

where

$$K^{1,*}_{j,l} = u^+_{j+\frac{1}{2}}\delta^{(l)}_{-1,j}(x_{j+\frac{1}{2}}) - u^+_{j-\frac{1}{2}}\delta^{(l)}_{-1,j}(x_{j-\frac{1}{2}}),\ K^{2,*}_{j,l} = q^-_{j+\frac{1}{2}}\delta^{(l)}_{-1,j}(x_{j+\frac{1}{2}}) - q^-_{j-\frac{1}{2}}\delta^{(l)}_{-1,j}(x_{j-\frac{1}{2}}),$$
$$K^{3,*}_{j,l} = (f, \delta^{(l)}_{-1,j}(x))_j, \qquad K^{4,*}_{j,l,l+1} = K^{1,*}_{j,l} - \tau K^{2,*}_{j,l+1} - K^{3,*}_{j,l+1}.$$

It is only $\bar{q}_j$ now that we do not know. Take $w = 1$ in (2.5), then we get $\bar{q}_j$ easily.

$$\bar{q}_j = \frac{1}{h}\left(u^+_{j+\frac{1}{2}} - u^+_{j-\frac{1}{2}}\right). \tag{B.6}$$

Collect the coefficients of $u^+_{j\pm\frac{1}{2}}$ and $q^-_{j\pm\frac{1}{2}}$ in the definitions of $K^4$ and $K^{4,*}$, then we have

$$\begin{aligned}\sum_{l=0}^{\frac{k-3}{2}} \tau^l K^4_{j,2l,2l+1} &= \left(\sum_{l=0}^{\frac{k-3}{2}} \tau^l \delta^{(2l)}_{1,j}(x_{j+\frac{1}{2}})\right) u^+_{j+\frac{1}{2}} - \left(\sum_{l=0}^{\frac{k-3}{2}} \tau^l \delta^{(2l)}_{1,j}(x_{j-\frac{1}{2}})\right) u^+_{j-\frac{1}{2}} \\ &- \left(\sum_{l=0}^{\frac{k-3}{2}} \tau^{l+1} \delta^{(2l+1)}_{1,j}(x_{j+\frac{1}{2}})\right) q^-_{j+\frac{1}{2}} + \left(\sum_{l=0}^{\frac{k-3}{2}} \tau^{l+1} \delta^{(2l+1)}_{1,j}(x_{j-\frac{1}{2}})\right) q^-_{j-\frac{1}{2}} - \sum_{l=0}^{\frac{k-3}{2}} \tau^l (f, \delta^{(2l+1)}_{1,j}(x))_j,\end{aligned} \tag{B.7}$$

and

$$\begin{aligned}\sum_{l=0}^{\frac{k-3}{2}} \tau^l K^{4,*}_{j,2l,2l+1} &= \left(\sum_{l=0}^{\frac{k-3}{2}} \tau^{l+1} \delta^{(2l+1)}_{-1,j}(x_{j+\frac{1}{2}})\right) u^+_{j+\frac{1}{2}} - \left(\sum_{l=0}^{\frac{k-3}{2}} \tau^{l+1} \delta^{(2l+1)}_{-1,j}(x_{j-\frac{1}{2}})\right) u^+_{j-\frac{1}{2}} \\ &- \left(\sum_{l=0}^{\frac{k-3}{2}} \tau^{l+2} \delta^{(2l+2)}_{-1,j}(x_{j+\frac{1}{2}})\right) q^-_{j+\frac{1}{2}} + \left(\sum_{l=0}^{\frac{k-3}{2}} \tau^{l+2} \delta^{(2l+2)}_{-1,j}(x_{j-\frac{1}{2}})\right) q^-_{j-\frac{1}{2}} - \sum_{l=0}^{\frac{k-3}{2}} \tau^{l+1} (f, \delta^{(2l+2)}_{-1,j}(x))_j.\end{aligned} \tag{B.8}$$

Put (B.7) and (B.8) into (B.4) and (B.5), we have

$$q^-_{j+\frac{1}{2}} = -\tau^{\frac{k-1}{2}} h\delta^{(k)}_{1,j}\bar{u}_j + \left(\sum_{l=0}^{\frac{k-1}{2}} \tau^l \delta^{(2l)}_{1,j}(x_{j+\frac{1}{2}})\right) u^+_{j+\frac{1}{2}} - \left(\sum_{l=0}^{\frac{k-1}{2}} \tau^l \delta^{(2l)}_{1,j}(x_{j-\frac{1}{2}})\right) u^+_{j-\frac{1}{2}}$$

(B.9)

$$- \left(\sum_{l=0}^{\frac{k-3}{2}} \tau^{l+1} \delta^{(2l+1)}_{1,j}(x_{j+\frac{1}{2}})\right) q^-_{j+\frac{1}{2}} + \left(\sum_{l=0}^{\frac{k-3}{2}} \tau^{l+1} \delta^{(2l+1)}_{1,j}(x_{j-\frac{1}{2}})\right) q^-_{j-\frac{1}{2}} - \sum_{l=0}^{\frac{k-3}{2}} \tau^l (f, \delta^{(2l+1)}_{1,j}(x))_j,$$

(B.10)

$$u^+_{j-\frac{1}{2}} = -\tau^{\frac{k+1}{2}} h\delta^{(k)}_{-1,j}\bar{q}_j + f^+_{j-\frac{1}{2}} + \tau\left(\sum_{l=0}^{\frac{k-1}{2}} \tau^l \delta^{(2l)}_{-1,j}(x_{j+\frac{1}{2}})\right) q^-_{j+\frac{1}{2}} - \tau\left(\sum_{l=0}^{\frac{k-1}{2}} \tau^l \delta^{(2l)}_{-1,j}(x_{j-\frac{1}{2}})\right) q^-_{j-\frac{1}{2}}$$
$$+ \sum_{l=1}^{\frac{k-1}{2}} \tau^l (f, \delta^{(2l)}_{-1,j}(x))_j - \left(\sum_{l=0}^{\frac{k-3}{2}} \tau^{l+1} \delta^{(2l+1)}_{-1,j}(x_{j+\frac{1}{2}})\right) u^+_{j+\frac{1}{2}} + \left(\sum_{l=0}^{\frac{k-3}{2}} \tau^{l+1} \delta^{(2l+1)}_{-1,j}(x_{j-\frac{1}{2}})\right) u^+_{j-\frac{1}{2}}$$

Use the first property of Lemma A.1, we get the following matrix form of (B.9) and (B.10)

$$\mathbf{D}\mathbf{q}^- = -\frac{1}{h}\sigma^{\frac{k-1}{2}} \frac{(2k+1)!}{k!}\bar{\mathbf{u}} + \frac{2}{h}\mathbf{B}\mathbf{u}^+ - \mathbf{f}_1, \tag{B.11a}$$

$$\mathbf{D}^\top \mathbf{u}^+ = \sigma^{\frac{k-1}{2}} \frac{\tau}{h} \frac{(2k+1)!}{k!}\bar{\mathbf{q}} - \frac{2\tau}{h}\mathbf{B}^\top \mathbf{q}^- + \mathbf{f}^+ + \mathbf{f}_2. \tag{B.11b}$$

By (B.3), (B.11a), (B.11b) and (B.6) and thanks to the commutability of circulant matrices with respect to matrix multiplication, we get

$$\mathbf{T}_3\bar{\mathbf{u}} = \mathbf{T}_4\mathbf{u}^+ - \frac{\tau}{h}\mathbf{A}\mathbf{f}_1 + \mathbf{D}\bar{\mathbf{f}}, \tag{B.12a}$$

$$\mathbf{T}_1\mathbf{u}^+ = \mathbf{T}_2\bar{\mathbf{u}} + \frac{2\tau}{h}\mathbf{f}_1 + \mathbf{D}(\mathbf{f}^+ + \mathbf{f}_2) \tag{B.12b}$$

Then put (B.12b) into (B.12a), and we have our final equation:

$$\mathbf{D}^{-1}(\mathbf{T}_1\mathbf{T}_3 - \mathbf{T}_2\mathbf{T}_4)\bar{\mathbf{u}} = \frac{\tau}{h}\mathbf{D}^{-1}\left(2\mathbf{T}_4\mathbf{B}^\top - \mathbf{T}_1\mathbf{A}\right)\mathbf{f}_1 + \mathbf{T}_1\bar{\mathbf{f}} + \mathbf{T}_4(\mathbf{f}^+ + \mathbf{f}_2).$$

□

For the convenience of analysis, we use point values to express the integral terms $\mathbf{f}_1$, $\mathbf{f}_2$ and $\bar{\mathbf{f}}$. These terms are integrals of polynomials with degree up to $2k-1$, therefore we can use LGL quadrature rule with $k+1$ points.

**Lemma B.2.** *Take $\{x_{j,\varrho}\}_{\varrho=0}^k$ as the LGL quadrature points in the subinterval $I_j$ and $\{w_\varrho\}_{\varrho=0}^k$ are the corresponding quadrature weights. Denote vector $\{\mathbf{f}^*_\varrho\}_{\varrho=0}^k \in \mathcal{R}^{N\times N}$ whose $j$-th component represents the function value of $f(x)$ at the $i$-th LGL numerical integration point corresponding to the $j$-th cell. Then the right vector in* (2.28) *can be*

*expressed by the following combination of point values* $\mathbf{f}_\varrho^*$.

$$\frac{\tau}{h}\mathbf{D}^{-1}\Big(2\mathbf{T}_4\mathbf{B}^\top-\mathbf{T}_1\mathbf{A}\Big)\mathbf{f}_1+\mathbf{T}_1\bar{\mathbf{f}}+\mathbf{T}_4(\mathbf{f}^+ +\mathbf{f}_2)=\sum_{\varrho=0}^{k}w_\varrho\mathbf{R}_\varrho\mathbf{f}_\varrho^*+\mathbf{T}_4\mathbf{f}^+, \tag{B.13}$$

*where* $\mathbf{R}_\varrho$ *is defined in Theorem 2.4 for each* $\varrho=0,\cdots,k$.

**Proof:** Note that $\mathbf{f}_1$, $\mathbf{f}_2$ and $\mathbf{T}_4$ are not all positive. To prove the positivity property, we need to represent all the integrals by point values through the LGL quadrature, as the integrated functions are all polynomials.

$$\frac{\tau}{h}\mathbf{D}^{-1}\Big(2\mathbf{T}_4\mathbf{B}^\top-\mathbf{T}_1\mathbf{A}\Big)\mathbf{f}_1+\mathbf{T}_1\bar{\mathbf{f}}+\mathbf{T}_4(\mathbf{f}^+ +\mathbf{f}_2).$$

Denote $\mathbf{P}=2\mathbf{T}_4\mathbf{B}^\top-\mathbf{T}_1\mathbf{A}$, then we have

$$\begin{aligned}\mathbf{P}&=2\Big(t_{41}\mathbf{I}+t_{42}\mathbf{G}+t_{43}\mathbf{G}^{N-1}\Big)\Big(-\gamma^*\mathbf{I}+\omega^*\mathbf{G}^{N-1}\Big)-\Big(t_{11}\mathbf{I}+t_{12}\mathbf{G}+t_{13}\mathbf{G}^{N-1}\Big)\Big(\mathbf{I}-\mathbf{G}^{N-1}\Big)\\&=\Big(-2t_{41}\gamma^*+2t_{42}\omega^*-t_{11}+t_{12}\Big)\mathbf{I}+\Big(-2\gamma^*t_{42}-t_{12}\Big)\mathbf{G}\\&+\Big(-2t_{43}\gamma^*+2t_{41}\omega^*+t_{11}-t_{13}\Big)\mathbf{G}^{N-1}+\Big(2\omega^*t_{43}+t_{13}\Big)\mathbf{G}^{N-2}\\&=\tilde{p}_1\mathbf{I}+\tilde{p}_2\mathbf{G}+\tilde{p}_3\mathbf{G}^{N-1}+\tilde{p}_4\mathbf{G}^{N-2},\end{aligned}$$

where

$$\begin{aligned}\tilde{p}_1&=-\Big((1+\xi)^2+\eta^2+(1+\xi)\eta+\sigma^{\frac{k+1}{2}}\frac{(2k+1)!}{k!}(2+2\xi+\eta)\Big),\\\tilde{p}_2&=(1+\xi)\eta+\sigma^{\frac{k+1}{2}}\frac{(2k+1)!}{k!}(1+\xi)=(1+\xi)\eta^*,\\\tilde{p}_3&=(1+\xi)^2+\eta^2+(1+\xi)\eta+\sigma^{\frac{k+1}{2}}\frac{(2k+1)!}{k!}(1+\xi+2\eta),\\\tilde{p}_4&=-\Big((1+\xi)\eta+\sigma^{\frac{k+1}{2}}\frac{(2k+1)!}{k!}\eta\Big)=-(1+\xi^*)\eta.\end{aligned}$$

Then we will show that the matrix $\mathbf{D}^{-1}\mathbf{P}$ is "periodic tridiagonal" which is similar with $\mathbf{D}^{-1}\mathbf{F}$. Notice that $\mathbf{D}^{-1}\mathbf{P}$ can be rewritten as a linear combination of $\{\mathbf{G}^{N-s}\}_{s=1}^N$. Firstly we prove that when $s=2,\cdots,N-2$, the coefficients of $\mathbf{G}^{N-s}$ are zero. Getting

$$\tilde{p}_1d^2+\tilde{p}_2d^3+\tilde{p}_3d+\tilde{p}_4=0$$

is enough. The left side of the equation above has two parts. The part which does not contain $\sigma$ is

$$\begin{aligned}&-\Big((1+\xi)^2+\eta^2+(1+\xi)\eta\Big)d^2+(1+\xi)\eta d^3+\Big((1+\xi)^2+\eta^2+(1+\xi)\eta\Big)d-(1+\xi)\eta\\=&(1+\xi)^2\Big(-d^2-d^4-d^3+d^4+d+d^3+d^2-d\Big)=0,\end{aligned}$$

and the other part is

$$-(2+2\xi+\eta)d^2+(1+\xi)d^3+(1+\xi+2\eta)d-\eta$$

$$= -2\frac{\eta^2}{1+\xi} - \frac{\eta^3}{(1+\xi)^2} + \frac{\eta^3}{(1+\xi)^2} + \eta + \frac{2\eta^2}{1+\xi} - \eta = 0.$$

Then the coefficients of $\mathbf{G}^{N-s}$ are zero for $s = 2, \cdots, N-2$. Then we have

$$\mathbf{D}^{-1}\mathbf{P} = \tilde{\tilde{p}}_1\mathbf{I} + \tilde{\tilde{p}}_2\mathbf{G} + \tilde{\tilde{p}}_3\mathbf{G}^{N-1},$$

where

$$\begin{aligned}
\tilde{\tilde{p}}_1 &= \frac{1}{(1+\xi)(1-d^N)}(\tilde{p}_1 + \tilde{p}_2 d + \tilde{p}_3 d^{N-1} + \tilde{p}_4 d^{N-2}) \\
&= \frac{1}{1+\xi}\left( -(1+\xi)^2 - (1+\xi)\eta - (2+2\xi)\sigma^{\frac{k+1}{2}}\frac{(2k+1)!}{k!}\right) = -1 - \xi^* - \eta^*, \\
\tilde{\tilde{p}}_2 &= \frac{1}{(1+\xi)(1-d^N)}\left((1+\xi)\eta(1-d^N) + (1+\xi)(1-d^N)\sigma^{\frac{k+1}{2}}\frac{(2k+1)!}{k!}\right) = \eta^*, \\
\tilde{\tilde{p}}_3 &= 1 + \xi^*.
\end{aligned}$$

Note that

$$\delta_{1,j}^{(2i+1)}(x) = \frac{2}{h}\left(\frac{2}{h}\right)^{2i+1}\delta_1^{(2i+1)}\left(\frac{2(x-x_j)}{h}\right) = \frac{2}{h}\left(\frac{2}{h}\right)^{2i+1}\delta_1^{(2i+1)}(\hat{x}),\ \hat{x} \in [-1,1],$$

and the degree of the integrated polynomials are at most $2k-1$. As a result, we can utilize the $k+1$ LGL point values to express the integral exactly. Therefore, we have

$$\frac{\tau}{h}\mathbf{f}_1 = \left(\frac{w_0}{2}\sum_{i=0}^{\frac{k-3}{2}}(4\sigma)^{i+1}\delta_1^{(2i+1)}(x_0)\right)\mathbf{f}_0^* + \cdots + \left(\frac{w_k}{2}\sum_{i=0}^{\frac{k-3}{2}}(4\sigma)^{i+1}\delta_1^{(2i+1)}(x_k)\right)\mathbf{f}_k^*, \tag{B.14}$$

and

$$\mathbf{f}_2 = \left(w_0\sum_{i=1}^{\frac{k-1}{2}}(4\sigma)^i\delta_{-1}^{(2i)}(x_0)\right)\mathbf{f}_0^* + \cdots + \left(w_k\sum_{i=1}^{\frac{k-1}{2}}(4\sigma)^i\delta_{-1}^{(2i)}(x_k)\right)\mathbf{f}_k^*. \tag{B.15}$$

Denote

$$f_{1,\varrho} = \frac{w_\varrho}{2}\sum_{i=0}^{\frac{k-3}{2}}(4\sigma)^{i+1}\delta_1^{(2i+1)}(x_\varrho),\ f_{2,\varrho} = w_\varrho\sum_{i=1}^{\frac{k-1}{2}}(4\sigma)^i\delta_{-1}^{(2i)}(x_\varrho), \tag{B.16}$$

and

$$\bar{f} = \frac{1}{h}\int_{I_j} f dx = \frac{1}{2}\frac{2}{h}\int_{I_j} f dx = \frac{1}{2}\sum_\varrho w_\varrho f_\varrho. \tag{B.17}$$

Then we get

$$\begin{aligned}
&\frac{\tau}{h}\mathbf{D}^{-1}\mathbf{P}\mathbf{f}_1 + \mathbf{T}_1\bar{\mathbf{f}} + \mathbf{T}_4(\mathbf{f}^+ + \mathbf{f}_2) \\
=&\mathbf{D}^{-1}\mathbf{P}\left(\sum_{\varrho=0}^{k} f_{1,\varrho}\mathbf{f}_\varrho\right) + \mathbf{T}_1\left(\frac{1}{2}\sum_{\varrho=0}^{k} w_\varrho\mathbf{f}_\varrho\right) + \mathbf{T}_4\left(\mathbf{f}^+ + \sum_{\varrho=0}^{k} f_{2,\varrho}\mathbf{f}_\varrho\right) \\
=&\sum_{\varrho=0}^{k}\left(f_{1,\varrho}\mathbf{D}^{-1}\mathbf{P} + \frac{1}{2}w_\varrho\mathbf{T}_1 + f_{2,\varrho}\mathbf{T}_4\right)\mathbf{f}_\varrho^* + \mathbf{T}_4\mathbf{f}^+.
\end{aligned}$$

□

B.2. **Proof of (i) in Theorem 2.5.**

**Proof:** We compute each element of $\mathbf{D}^{-1}(\mathbf{T}_1\mathbf{T}_3 - \mathbf{T}_2\mathbf{T}_4)$ to check whether it is an $M$-matrix. First we compute $\mathbf{T}_\varrho$, $(\varrho = 1, 2, 3, 4)$ with their definitions. For example: (B.18)

$$\begin{aligned}\mathbf{T}_1 =&\mathbf{D}^\top\mathbf{D} + \sigma^{\frac{k+1}{2}}\frac{(2k+1)!}{k!}\mathbf{A}^\top\mathbf{D} + 4\sigma\mathbf{B}^\top\mathbf{B}\\ =&[(1+\xi)\mathbf{I} - \eta\mathbf{G}^{N-1}][(1+\xi)\mathbf{I} - \eta\mathbf{G}] + \sigma^{\frac{k+1}{2}}\frac{(2k+1)!}{k!}(\mathbf{I}-\mathbf{G})[(1+\xi)\mathbf{I} - \eta\mathbf{G}^{N-1}]\\ &+ 4\sigma(-\gamma^*\mathbf{I} + \omega^*\mathbf{G}^{N-1})(-\gamma^*\mathbf{I} + \omega^*\mathbf{G}) = t_{11}\mathbf{I} + t_{12}\mathbf{G} + t_{13}\mathbf{G}^{N-1},\end{aligned}$$

where

$$\begin{aligned} t_{11} &= (1+\xi)^2 + \eta^2 + \sigma^{\frac{k+1}{2}}\frac{(2k+1)!}{k!}(1+\xi+\eta) + 4\sigma(\gamma^*)^2 + 4\sigma(\omega^*)^2,\\ t_{12} &= -\bigg((1+\xi)\eta + (1+\xi)\sigma^{\frac{k+1}{2}}\frac{(2k+1)!}{k!} + 4\sigma\gamma^*\omega^*\bigg),\\ t_{13} &= -\bigg((1+\xi)\eta + \eta\sigma^{\frac{k+1}{2}}\frac{(2k+1)!}{k!} + 4\sigma\gamma^*\omega^*\bigg).\end{aligned} \tag{B.19}$$

Similarly, we get

$$\mathbf{T}_i = t_{i1}\mathbf{I} + t_{i2}\mathbf{G} + t_{i3}\mathbf{G}^{N-1}, \qquad \mathrm{i} = 2, 3, 4,$$

where

$$\begin{aligned} &t_{21} = -2\sigma^{\frac{k+1}{2}}\frac{(2k+1)!}{k!}\gamma^*, && t_{22} = 0, && t_{23} = 2\sigma^{\frac{k+1}{2}}\frac{(2k+1)!}{k!}\omega^*,\\ &t_{31} = (1+\xi) + \sigma^{\frac{k+1}{2}}\frac{(2k+1)!}{k!}, && t_{32} = 0, && t_{33} = -\eta - \sigma^{\frac{k+1}{2}}\frac{(2k+1)!}{k!},\\ &t_{41} = -2\sigma(\gamma^* + \omega^*), && t_{42} = 2\sigma\omega^*, && t_{43} = 2\sigma\gamma^*.\end{aligned} \tag{B.20}$$

Denote $\mathbf{F} = \mathbf{T}_1\mathbf{T}_3 - \mathbf{T}_2\mathbf{T}_4$, then we get

$$\mathbf{F} = \tilde{f}_1\mathbf{I} + \tilde{f}_2\mathbf{G} + \tilde{f}_3\mathbf{G}^{N-1} + \tilde{f}_4\mathbf{G}^{N-2},$$

where

$$\begin{aligned} &\tilde{f}_1 = t_{11}t_{31} + t_{12}t_{33} - t_{41}t_{21} - t_{42}t_{23}, \quad \tilde{f}_2 = t_{31}t_{12} - t_{21}t_{42},\\ &\tilde{f}_3 = t_{31}t_{13} + t_{11}t_{33} - t_{43}t_{21} - t_{41}t_{23}, \quad \tilde{f}_4 = t_{13}t_{33} - t_{43}t_{23}.\end{aligned} \tag{B.21}$$

We take $t_{ij}$ in for more details

$$\begin{aligned}\tilde{f}_1 =&t_{11}t_{31} + t_{12}t_{33} - t_{41}t_{21} - t_{42}t_{23}\\ =&[(1+\xi)^3 + 2(1+\xi)\eta^2] + [4\sigma((\omega^*)^2 + (\gamma^*)^2)(1+\xi) + 4\sigma\gamma^*\omega^*\eta]\\ &+ [2(1+\xi)^2 + 3(1+\xi)\eta + \eta^2]\sigma^{\frac{k+1}{2}}\frac{(2k+1)!}{k!} + [2(1+\xi)+\eta]\bigg(\sigma^{\frac{k+1}{2}}\frac{(2k+1)!}{k!}\bigg)^2,\\ \tilde{f}_2 =&t_{31}t_{12} - t_{21}t_{42}\\ =&-(1+\xi)^2\eta - [(1+\xi)^2 + (1+\xi)\eta]\sigma^{\frac{k+1}{2}}\frac{(2k+1)!}{k!} - (1+\xi)\bigg(\sigma^{\frac{k+1}{2}}\frac{(2k+1)!}{k!}\bigg)^2 - 4\sigma\gamma^*\omega^*(1+\xi),\end{aligned}$$

(B.22)

$$\begin{aligned}\tilde{f}_3 =& t_{31}t_{13} + t_{11}t_{33} - t_{43}t_{21} - t_{41}t_{23}\\ =& [-2(1+\xi)^2\eta - \eta^3] - [(1+\xi)4\sigma\gamma^*\omega^* + \eta 4\sigma((\omega^*)^2 + (\gamma^*)^2)]\\ & - [3(1+\xi)\eta + (1+\xi)^2 + 2\eta^2]\sigma^{\frac{k+1}{2}}\frac{(2k+1)!}{k!} - (2\eta + 1 + \xi)\left(\sigma^{\frac{k+1}{2}}\frac{(2k+1)!}{k!}\right)^2,\\ \tilde{f}_4 =& t_{13}t_{33} - t_{43}t_{23}\\ =& (1+\xi)\eta^2 + 4\sigma\gamma^*\omega^*\eta + [(1+\xi)\eta + \eta^2]\sigma^{\frac{k+1}{2}}\frac{(2k+1)!}{k!} + \eta\left(\sigma^{\frac{k+1}{2}}\frac{(2k+1)!}{k!}\right)^2.\end{aligned}$$

Note that $\mathbf{D} = (1+\xi)\mathbf{I} - \eta\mathbf{G}^{N-1}$. As long as $1+\xi \neq \eta$ and $1+\xi \neq 0$, we always have

$$\mathbf{D}^{-1} = \frac{1}{(1+\xi)(1-d^N)}\sum_{s=0}^{N-1} d^s\mathbf{G}^{N-s},$$

where $d = \frac{\eta}{1+\xi}$. The above property can be checked directly. Then with (B.22), we have

(B.23)

$$\begin{aligned}\mathbf{D}^{-1}\mathbf{F} =& \frac{1}{(1+\xi)(1-d^N)}\sum_{s=0}^{N-1} d^s\mathbf{G}^{N-s}(\tilde{f}_1\mathbf{I} + \tilde{f}_2\mathbf{G} + \tilde{f}_3\mathbf{G}^{N-1} + \tilde{f}_4\mathbf{G}^{N-2})\\ =& \frac{1}{(1+\xi)(1-d^N)}(\tilde{f}_1\sum_{s=0}^{N-1} d^s + \tilde{f}_2\sum_{s=-1}^{N-2} d^{s+1} + \tilde{f}_3\sum_{s=1}^{N} d^{s-1} + \tilde{f}_4\sum_{s=2}^{N+1} d^{s-2})\mathbf{G}^{N-s}\\ =& \frac{1}{(1+\xi)(1-d^N)}\Bigg((\tilde{f}_1 + \tilde{f}_2 d + \tilde{f}_3 d^{N-1} + \tilde{f}_4 d^{N-2})\mathbf{I} + (\tilde{f}_1 d^{N-1} + \tilde{f}_2 + \tilde{f}_3 d^{N-2} + \tilde{f}_4 d^{N-3})\mathbf{G}\\ & + (\tilde{f}_1 d + \tilde{f}_2 d^2 + \tilde{f}_3 + \tilde{f}_4 d^{N-1})\mathbf{G}^{N-1} + \sum_{s=2}^{N-2}\frac{d^{s-2}}{(1+\xi)(1-d^N)}(\tilde{f}_1 d^2 + \tilde{f}_2 d^3 + \tilde{f}_3 d + \tilde{f}_4)\mathbf{G}^{N-s}\Bigg).\end{aligned}$$

By computing carefully, we can get

$$\tilde{f}_1 d^2 + \tilde{f}_2 d^3 + \tilde{f}_3 d + \tilde{f}_4 = 0.$$

Then we will prove the above equation. Denote $d = \frac{\eta}{1+\xi}$. We check carefully by combining similar terms. For the coefficients of the terms which have $\xi$ and $\eta$ only in $f_1 d^2 + f_2 d^3 + f_3 d + f_4$, we have

$$\begin{aligned}& \left((1+\xi)^3 + 2(1+\xi)\eta^2\right)d^2 - (1+\xi)^2\eta d^3 - \left(2(1+\xi)^2\eta + \eta^3\right)d + (1+\xi)\eta^2\\ =& (1+\xi)^3\left((1+2d^2)d^2 - d^4 - (2d + d^3)d + d^2\right) = 0.\end{aligned}$$

For the coefficients of the terms which have $\omega^*$ and $\gamma^*$, we have

$$\begin{aligned}& 4\sigma\left((\gamma^*)^2 + (\omega^*)^2\right)\left((1+\xi)(\frac{\eta}{1+\xi})^2 - \eta(\frac{\eta}{1+\xi})\right)\\ & + 4\sigma\gamma^*\omega^*\left(\eta(\frac{\eta}{1+\xi})^2 - (1+\xi)(\frac{\eta}{1+\xi})^3 - (1+\xi)\frac{\eta}{1+\xi} + \eta\right) = 0.\end{aligned}$$

For the coefficients of the terms which have $\sigma^{\frac{k+1}{2}}\frac{(2k+1)!}{k!}$, we have

$$\begin{aligned}&\Big(2(1+\xi)^2+3(1+\xi)\eta+\eta^2\Big)d^2-\Big((1+\xi)^2+(1+\xi)\eta\Big)d^3\\&-\Big(3(1+\xi)\eta+(1+\xi)^2+2\eta^2\Big)d+(1+\xi)\eta+\eta^2\\=&(1+\xi)^2\Big((2+3d+d^2)d^2-(1+d)d^3-(3d+1+2d^2)d+d+d^2\Big)=0.\end{aligned}$$

And for the coefficients of the terms which have $\left(\sigma^{\frac{k+1}{2}}\frac{(2k+1)!}{k!}\right)^2$, we have

$$\begin{aligned}&\left(2(1+\xi)+\eta\right)d^2-(1+\xi)d^3-(2\eta+(1+\xi))d+\eta\\&=(1+\xi)\Big((2+d)d^2-d^3-(2d+1)d+d\Big)=0.\end{aligned}$$

Then we have $\tilde{f}_1d^2+\tilde{f}_2d^3+\tilde{f}_3d+\tilde{f}_4=0$. Hence $\mathbf{D}^{-1}\mathbf{F}$ can be written as

(B.24)

$$\begin{aligned}\mathbf{D}^{-1}\mathbf{F}=&\frac{1}{(1+\xi)(1-d^N)}\Big((\tilde{f}_1+\tilde{f}_2d+\tilde{f}_3d^{N-1}+\tilde{f}_4d^{N-2})\mathbf{I}+(\tilde{f}_1d^{N-1}+\tilde{f}_2+\tilde{f}_3d^{N-2}+\tilde{f}_4d^{N-3})\mathbf{G}\\&+(\tilde{f}_1d+\tilde{f}_2d^2+\tilde{f}_3+\tilde{f}_4d^{N-1})\mathbf{G}^{N-1}\Big)\\=&\tilde{\tilde{f}}_1\mathbf{I}+\tilde{\tilde{f}}_2\mathbf{G}+\tilde{\tilde{f}}_3\mathbf{G}^{N-1},\end{aligned}$$

where

$$\tilde{\tilde{f}}_1=(1+\xi)^2(1+d^2)+4\sigma\Big((\gamma^*)^2+(\omega^*)^2\Big)+2(1+\xi)(1+d)\sigma^{\frac{k+1}{2}}\frac{(2k+1)!}{k!}+2\left(\sigma^{\frac{k+1}{2}}\frac{(2k+1)!}{k!}\right)^2,$$

$$\tilde{\tilde{f}}_2=-d(1+\xi)^2-4\sigma\gamma^*\omega^*-(1+d)(1+\xi)\sigma^{\frac{k+1}{2}}\frac{(2k+1)!}{k!}-\left(\sigma^{\frac{k+1}{2}}\frac{(2k+1)!}{k!}\right)^2,$$

$$\tilde{\tilde{f}}_3=\tilde{\tilde{f}}_2.$$

Then notice that the left matrix of (2.28) have the form as below

$$\mathbf{D}^{-1}\mathbf{F}=\begin{pmatrix}\tilde{\tilde{f}}_1&\tilde{\tilde{f}}_2&0&\cdots&\tilde{\tilde{f}}_2\\\tilde{\tilde{f}}_2&\tilde{\tilde{f}}_1&\tilde{\tilde{f}}_2&\cdots&0\\0&\tilde{\tilde{f}}_2&\ddots&\ddots&\vdots\\\vdots&\ddots&\ddots&\ddots&\tilde{\tilde{f}}_2\\\tilde{\tilde{f}}_2&0&\cdots&\tilde{\tilde{f}}_2&\tilde{\tilde{f}}_1\end{pmatrix}.$$

To prove $\mathbf{D}^{-1}\mathbf{F}$ is an $M$-matrix, we only need to prove $\tilde{\tilde{f}}_2<0$ and $\tilde{\tilde{f}}_1>-2|\tilde{\tilde{f}}_2|$, which means $\mathbf{D}^{-1}\mathbf{F}$ is diagonally dominant. Notice that

$$\begin{aligned}\tilde{\tilde{f}}_1+\tilde{\tilde{f}}_2+\tilde{\tilde{f}}_3&=(1+\xi)^2(1-\frac{\eta}{1+\xi})^2+4\sigma(\gamma^*-\omega^*)^2\\&=(1+\xi-\eta)^2+4\sigma(\gamma^*-\omega^*)^2>0,\end{aligned}$$

and

$$\tilde{\tilde{f}}_1 - \tilde{\tilde{f}}_2 - \tilde{\tilde{f}}_3 = (1+\xi)^2(1+d)^2 + 4\sigma(\gamma^*+\omega^*)^2 + 4(\sigma\frac{(2k+1)!}{k!})^2 + 4(1+\xi)(1+d)\sigma^{\frac{k+1}{2}}\frac{(2k+1)!}{k!}$$
$$= \left(2\sigma^{\frac{k+1}{2}} + (1+\xi)(1+d)\right)^2 + 4\sigma(\gamma^*+\omega^*)^2 > 0.$$

As a result, we have

$$\tilde{\tilde{f}}_1 > |\tilde{\tilde{f}}_2 + \tilde{\tilde{f}}_3| = |\tilde{\tilde{f}}_2| + |\tilde{\tilde{f}}_3|.$$

Next we only need to prove $\tilde{\tilde{f}}_2 < 0$. Thanks to

$$(4\sigma)^{\frac{k+1}{2}}\delta_1^{(k)}(x) = \frac{(4\sigma)^{\frac{k+1}{2}}(2k+1)}{2}p_k^{(k)}(1) = \frac{(4\sigma)^{\frac{k+1}{2}}(2k+1)}{2}\frac{(2k)!}{2^k k!} = (\sigma)^{\frac{k+1}{2}}\frac{(2k+1)!}{k!},$$

we have

$$\begin{aligned}\tilde{\tilde{f}}_2 &= -d(1+\xi)^2 - 4\sigma\gamma^*\omega^* - (1+d)(1+\xi)\sigma^{\frac{k+1}{2}}\frac{(2k+1)!}{k!} - \left(\sigma^{\frac{k+1}{2}}\frac{(2k+1)!}{k!}\right)^2 \\ &= -\left(\sigma^{\frac{k+1}{2}}\frac{(2k+1)!}{k!} + 1 + \xi\right)\left(\sigma^{\frac{k+1}{2}}\frac{(2k+1)!}{k!} + d(1+\xi)\right) - 4\sigma\gamma^*\omega^* \\ &= -(1+\xi^*)\eta^* - 4\sigma\gamma^*\omega^*.\end{aligned}$$

To get $\tilde{\tilde{f}}_2 < 0$ and finish this lemma, we need to prove

$$(1+\xi^*)\eta^* + 4\sigma\gamma^*\omega^* > 0.$$

By definition, we have

$$\xi^* = \sum_{i=0}^{\frac{k-1}{2}}(4\sigma)^{i+1}\delta_1^{(2i+1)}(1), \qquad \eta^* = \sum_{i=0}^{\frac{k-1}{2}}(4\sigma)^{i+1}\delta_1^{(2i+1)}(-1),$$
$$\omega^* = \sum_{i=0}^{\frac{k-1}{2}}(4\sigma)^{i}\delta_1^{(2i)}(1), \qquad \gamma^* = \sum_{i=0}^{\frac{k-1}{2}}(4\sigma)^{i}\delta_1^{(2i)}(-1).$$

It is easy to see (or by Lemma A.2) that

$$\xi^* > \eta^* > 0 > \gamma^*,\ \omega^* \geq |\gamma^*| > 0.$$

Then we have

$$\begin{aligned}&(1+\xi^*)\eta^* + 4\sigma\gamma^*\omega^* \\ =&\sum_{i_1=0}^{\frac{k-1}{2}}\sum_{i_2=0}^{\frac{k-1}{2}}(4\sigma)^{i_1+i_2+2}\delta_1^{(2i_1+1)}(1)\delta_1^{(2i_2+1)}(-1) + \sum_{i_1=0}^{\frac{k-1}{2}}\sum_{i_2=0}^{\frac{k-1}{2}}(4\sigma)^{i_1+i_2+1}\delta_1^{(2i_1)}(1)\delta_1^{(2i_2)}(-1) \\ =&\sum_{i_1=0}^{\frac{k-1}{2}}\sum_{i_2=0}^{\frac{k-1}{2}}(4\sigma)^{i_1+i_2+1}\left(4\sigma\delta_1^{(2i_1+1)}(1)\delta_1^{(2i_2+1)}(-1) + \delta_1^{(2i_1)}(1)\delta_1^{(2i_2)}(-1)\right).\end{aligned}$$

Using the third and the fourth properties of Lemma A.2, we have

(B.26)

$$\begin{aligned}&4\sigma\delta_1^{(2i_1+1)}(1)\delta_1^{(2i_2+1)}(-1)+\delta_1^{(2i_1)}(1)\delta_1^{(2i_2)}(-1)\\=&4\sigma\delta_1^{(2i_1+1)}(1)\Bigg(\underbrace{\delta_1^{(2i_2+1)}(-1)+\delta_1^{(2i_2)}(-1)}_{>0}\Bigg)-\underbrace{\delta_1^{(2i_2)}(-1)}_{<0}\Bigg(4\sigma\delta_1^{(2i_1+1)}(1)-\delta_1^{(2i_1)}(1)\Bigg).\end{aligned}$$

Next we want to show $4\sigma\delta_1^{(2i+1)}(1)-\delta_1^{(2i)}(1)>0$. Notice that

$$\begin{aligned}&2\delta_1^{(2i+1)}(1)-\delta_1^{(2i)}(1)=\frac{2}{c_{2i+1}}\sum_{l=2i+1}^{k}\nu_{2i+1}^l-\frac{1}{c_{2i}}\sum_{l=2i}^{k}\nu_{2i}^l\\=&\frac{1}{c_{2i}}\sum_{l=2i+1}^{k}\left(\frac{1}{2i+1}\nu_{2i+1}^l-\nu_{2i}^l\right)-\frac{1}{c_{2i}}\nu_{2i}^{2i},\end{aligned}$$

where $\delta_1^{(i)}(1)=\frac{1}{c_i}\sum_{l=i}^{k}\nu_i^l$ and $c_i=2^{i+1}i!$, $\nu_i^l=(2l+1)\frac{(l+i)!}{(l-i)!}$. Also we find

$$\begin{aligned}&\frac{1}{2i+1}\nu_{2i+1}^l-\nu_{2i}^l=\frac{1}{2i+1}(2l+1)\frac{(l+2i+1)!}{(l-2i-1)!}-(2l+1)\frac{(l+2i)!}{(l-2i)!}\\=&(\frac{l+2i+1}{2i+1})(2l+1)\frac{(l+2i)!}{(l-2i-1)!}-(2l+1)\frac{(l+2i)!}{(l-2i)!}>0\end{aligned}$$

holds for any $i=0,\cdots,\frac{k-1}{2},\ l=2i,\cdots,k$, and

$$\frac{1}{2i+1}\nu_{2i+1}^k-\nu_{2i}^k-\nu_{2i}^{2i}=\underbrace{(2k+1)}_{\geq 4i+1}(\underbrace{\frac{k+2i+1}{2i+1}}_{\geq 2}-\underbrace{\frac{1}{k-2i}}_{\leq 1})\frac{(k+2i)!}{(k-2i-1)!}-(4i+1)\frac{(3i)!}{i!}>0$$

holds due to an observation that each term of $\frac{(k+2i)!}{(k-2i-1)!}$ is larger than the corresponding term of $\frac{(3i)!}{i!}$. Then we get

$$\begin{aligned}&2\delta_1^{(2i+1)}(1)-\delta_1^{(2i)}(1)=\frac{2}{c_{2i+1}}\sum_{l=2i+1}^{k}\nu_{2i+1}^l-\frac{1}{c_{2i}}\sum_{l=2i}^{k}\nu_{2i}^l\\=&\frac{1}{c_{2i}}\sum_{l=2i+1}^{k-1}\left(\frac{1}{2i+1}\nu_{2i+1}^l-\nu_{2i}^l\right)+\frac{1}{c_{2i}}\left(\frac{1}{2i+1}\nu_{2i+1}^k-\nu_{2i}^k-\nu_{2i}^{2i}\right)>0.\end{aligned}$$

It is known that $\delta_1^{(2i+1)}(1)>0$. As a result, if $\sigma\geq\frac{1}{2}$, we have

$$4\sigma\delta_1^{(2i+1)}(1)-\delta_1^{(2i)}(1)>2\delta_1^{(2i+1)}(1)-\delta_1^{(2i)}(1)>0.$$

Then we have $\tilde{\tilde{f}}_2<0$. Furthermore, we get $\mathbf{D}^{-1}\mathbf{F}$ is an $M$-matrix. □

B.3. **Proof of (ii) in Theorem 2.5.** Recall that

$$\mathbf{R}_\varrho=f_{1,\varrho}\mathbf{D}^{-1}\mathbf{P}+\frac{1}{2}w_\varrho\mathbf{T}_1+f_{2,\varrho}\mathbf{T}_4,\ \varrho=0,\cdots,k.$$

We have

$$\begin{aligned}\mathbf{R}_\varrho =&f_{1,\varrho}\mathbf{D}^{-1}\mathbf{P}+\frac{1}{2}w_\varrho\mathbf{T}_1+f_{2,\varrho}\mathbf{T}_4\\ =&f_{1,\varrho}(\tilde{p}_1\mathbf{I}+\tilde{p}_2\mathbf{G}+\tilde{p}_3\mathbf{G}^{N-1})+\frac{w_\varrho}{2}(t_{11}\mathbf{I}+t_{12}\mathbf{G}+t_{13}\mathbf{G}^{N-1})+f_{2,\varrho}(t_{41}\mathbf{I}+t_{42}\mathbf{G}+t_{43}\mathbf{G}^{N-1})\\ =&(f_{1,\varrho}\tilde{p}_1+\frac{w_j}{2}t_{11}+f_{2,\varrho}t_{41})\mathbf{I}+(f_{1,\varrho}\tilde{p}_2+\frac{w_\varrho}{2}t_{12}+f_{2,\varrho}t_{42})\mathbf{G}+(f_{1,\varrho}\tilde{p}_3+\frac{w_\varrho}{2}t_{13}+f_{2,\varrho}t_{43})\mathbf{G}^{N-1}\\ =&r_{1,\varrho}\mathbf{I}+r_{2,\varrho}\mathbf{G}+r_{3,\varrho}\mathbf{G}^{N-1},\end{aligned}\tag{B.27}$$

where $f_{i',\varrho}$ $i'=1,2,3$, are defined in (B.16). Note that

$$\delta_{-1}^{(2i)}(x)=\sum_{l=0}^{k}\frac{2l+1}{2}(-1)^l p_l^{(2i)}(x)=\sum_{l=0}^{k}\frac{2l+1}{2}(-1)^{l+2i}p_l^{(2i)}(x)=\sum_{l=0}^{k}\frac{2l+1}{2}p_l^{(2i)}(-x)=\delta_1^{(2i)}(-x),$$

then denote

$$r_1(\sigma,x)=\frac{1}{2}\sum_{i=0}^{\frac{k-3}{2}}(4\sigma)^{i+1}\delta_1^{(2i+1)}(x)(-1-\xi^*-\eta^*)+\frac{1}{2}\bigg((1+\xi)^2+\eta^2+\sigma^{\frac{k+1}{2}}\frac{(2k+1)!}{k!}(1+\xi+\eta)\bigg)\tag{B.28a}$$

$$+\frac{1}{2}\bigg(4\sigma((\gamma^*)^2+(\omega^*)^2)\bigg)-\frac{1}{2}\bigg(4\sigma(\gamma^*+\omega^*)\bigg)\sum_{i=1}^{\frac{k-1}{2}}(4\sigma)^i\delta_1^{(2i)}(-x),$$

$$r_2(\sigma,x)=\frac{1}{2}\sum_{i=0}^{\frac{k-3}{2}}(4\sigma)^{i+1}\delta_1^{(2i+1)}(x)(\eta^*)-\frac{1}{2}\bigg((1+\xi)\eta^*+4\sigma\gamma^*\omega^*\bigg)+\frac{1}{2}\sum_{i=1}^{\frac{k-1}{2}}(4\sigma)^i\delta_1^{(2i)}(-x)(4\sigma\omega^*),\tag{B.28b}$$

$$r_3(\sigma,x)=\frac{1}{2}\bigg(\sum_{i=1}^{\frac{k-1}{2}}(4\sigma)^i\delta_1^{(2i-1)}(x)-\eta\bigg)(1+\xi^*)+2\sigma\gamma^*\bigg(\sum_{i=1}^{\frac{k-1}{2}}(4\sigma)^i\delta_1^{(2i)}(-x)-\omega^*\bigg).\tag{B.28c}$$

We will prove that $r_1(\sigma,x)$, $r_2(\sigma,x)$, and $r_3(\sigma,x)$, instead of $r_{1,\varrho},r_{2,\varrho},r_{3,\varrho}$, are non-negative for fixed $x$ on $(-1,1)$ when $\sigma$ has a lower bound. Meanwhile, we shall also prove additionally the coefficient matrices of $\mathbf{f}^+$ and $\mathbf{f}^-$ are non-negative. Therefore, we will get the second property of Theorem 2.5. Firstly, we prove that $r_1(\sigma,x)>0$ when $\sigma\geq\frac{1}{2}$ and $k\geq 3$ and we begin by introducing the following lemma.

**Lemma B.3.** *Define*

$$F(\sigma,x):=\sum_{i=0}^{\frac{k-1}{2}}(4\sigma)^i\delta_1^{(2i)}(x).$$

*If $\sigma\geq\frac{1}{2}$ and $k\geq 3$, then the $p$-th derivative of $F(\sigma,x)$ is increasing monotonically when $p$ is even and is positive when $p$ is odd, which means that*

$$\frac{\partial^{2i_1}F(\sigma,x)}{\partial x^{2i_1}}>\frac{\partial^{2i_1}F(\sigma,-1)}{\partial x^{2i_1}},\tag{B.29a}$$

$$\frac{\partial^{2i_1+1}F(\sigma,x)}{\partial x^{2i_1+1}}>0, \tag{B.29b}$$

*where* $i_1=0,1,\cdots,\frac{k-1}{2}$.

**Proof:** It is clear that (B.29b) implies (B.29a). We now prove (B.29a). Note that

$$\frac{\partial^k}{\partial x^k}F(\sigma,x)=\delta_1^{(k)}(x)=\frac{2k+1}{2}p_k^{(k)}(x)>0.$$

Suppose we already know

$$\frac{\partial^{k-2i_1}}{\partial x^{k-2i_1}}F(\sigma,x)>0,$$

then, for using recurrence method, we shall prove that

$$\frac{\partial^{k-2i_1-2}}{\partial x^{k-2i_1-2}}F(\sigma,x)>0,$$

where $i_1=0,1,\cdots,\frac{k-3}{2}$. Note that $\frac{\partial^{k-2i_1-1}}{\partial x^{k-2i_1-1}}F(\sigma,1)>0>\frac{\partial^{k-2i_1-1}}{\partial x^{k-2i_1-1}}F(\sigma,-1)$ (the fourth property in Lemma A.2), so the function $\frac{\partial^{k-2i_1-2}}{\partial x^{k-2i_1-2}}F(\sigma,x)$ has a minimum point which is denoted as $x_{i_1,0}$. By Lemma A.1, we have

$$p_k^{(k)}\geq p_k^{(k-i_1)}(1)i_1!\geq|p_k^{(k-i_1)}(x)|i_1!. \tag{B.30}$$

Furthermore, we have

$$\begin{aligned}
&\frac{\partial^{k-2i_1-2}}{\partial x^{k-2i_1-2}}F(\sigma,x)\geq\frac{\partial^{k-2i_1-2}}{\partial x^{k-2i_1-2}}F(\sigma,x_{i_1,0})=\sum_{i=0}^{\frac{k-1}{2}}(4\sigma)^i\sum_{l=0}^{k}\frac{2l+1}{2}p_l^{(2i+k-2i_1-2)}(x_{i_1,0})\\
=&\sum_{i=0}^{i_1+1}(4\sigma)^i\sum_{l=2i+k-2i_1-2}^{k}\frac{2l+1}{2}p_l^{(2i+k-2i_1-2)}(x_{i_1,0})\\
=&\sum_{i=0}^{i_1}(4\sigma)^i\left(\sum_{l=2i+k-2i_1-2}^{k-1}\frac{2l+1}{2}p_l^{(2i+k-2i_1-2)}(x_{i_1,0})\right)+\frac{2k+1}{2}\left(\sum_{i=0}^{i_1}(4\sigma)^ip_k^{(2i+k-2i_1-2)}(x_{i_1,0})+(4\sigma)^{i_1+1}p_k^{(k)}\right)\\
=&\sum_{i=0}^{i_1}(4\sigma)^i\sum_{l=k-1}^{k}\frac{1}{2}p_l^{(2i+k-2i_1-1)}(x_{i_1,0})+\frac{2k+1}{2}\left(\sum_{i=0}^{i_1}(4\sigma)^ip_k^{(2i+k-2i_1-2)}(x_{i_1,0})+(4\sigma)^{i_1+1}p_k^{(k)}\right)\\
=&\sum_{i=0}^{i_1}(4\sigma)^i\sum_{l=k-1}^{k}\frac{1}{2}\left(p_l^{(2i+k-2i_1-1)}(x_{i_1,0})+\frac{1}{(2i_1+1-2i)!}p_k^{(k)}\right)\\
&+\frac{2k+1}{2}\left(\sum_{i=0}^{i_1}(4\sigma)^i\left(p_k^{(2i+k-2i_1-2)}(x_{i_1,0})+\frac{1}{(2i_1+2-2i)!}p_k^{(k)}\right)\right)\\
&+\left((4\sigma)^{i_1+1}\frac{2k+1}{2}-\sum_{i=0}^{i_1}(4\sigma)^i\frac{1}{(2i_1+1-2i)!}-\frac{2k+1}{2}\sum_{i=0}^{i_1}(4\sigma)^i\frac{1}{(2i_1+2-2i)!}\right)p_k^{(k)}.
\end{aligned}$$

Then we only need to prove, for any $i_1 = 0, \cdots, \frac{k-1}{2}$, the following inequality always holds

$$(4\sigma)^{i_1+1}\frac{2k+1}{2} - \sum_{i=0}^{i_1}(4\sigma)^i \frac{1}{(2i_1+1-2i)!} - \frac{2k+1}{2}\sum_{i=0}^{i_1}(4\sigma)^i\frac{1}{(2i_1+2-2i)!} > 0.$$

It is clear that the above condition holds as long as $\sigma \geq \frac{1}{2}$ and $k \geq 3$. □

Then by the above lemma, we introduce the following theorem.

**Theorem B.1.** *When $\sigma \geq \frac{1}{2}$ we have*

$$r_1(\sigma, x) > 0, \ x \in [-1, 1].$$

**Proof:** By (B.28a) and the above lemma, we have

$$\begin{aligned}
r_1(\sigma,x) &= \frac{1}{2}\sum_{i=0}^{\frac{k-3}{2}}(4\sigma)^{i+1}\delta_1^{(2i+1)}(x)(-1-\xi^*-\eta^*) + \frac{1}{2}\bigg((1+\xi)^2+\eta^2+\sigma^{\frac{k+1}{2}}\frac{(2k+1)!}{k!}(1+\xi+\eta)\bigg) \\
&+ \frac{1}{2}\bigg(4\sigma((\gamma^*)^2+(\omega^*)^2)\bigg) - \frac{1}{2}\bigg(4\sigma(\gamma^*+\omega^*)\bigg)\sum_{i=1}^{\frac{k-1}{2}}(4\sigma)^i\delta_1^{(2i)}(-x) \\
=&\frac{1}{2}\bigg(-\sum_{i=0}^{\frac{k-3}{2}}(4\sigma)^{i+1}\delta_1^{(2i+1)}(x)+1+\xi\bigg)(1+\xi^*) + 2\sigma\gamma^*\bigg(\sum_{i=0}^{\frac{k-1}{2}}(4\sigma)^i\delta_1^{(2i)}(-1) - \sum_{i=1}^{\frac{k-1}{2}}(4\sigma)^{i+1}\delta_1^{(2i+1)}(-x)\bigg) \\
&+ \frac{1}{2}\eta^*\bigg(-\sum_{i=0}^{\frac{k-3}{2}}(4\sigma)^{i+1}\delta_1^{(2i+1)}(x)+\eta\bigg) + 2\sigma\omega^*\bigg(\omega^* - \sum_{i=1}^{\frac{k-1}{2}}(4\sigma)^i\delta_1^{(2i)}(-x)\bigg) \\
>&\frac{1}{2}\eta^*\bigg(-\sum_{i=0}^{\frac{k-3}{2}}(4\sigma)^{i+1}\delta_1^{(2i+1)}(x)+\eta\bigg) + 2\sigma\omega^*\bigg(\omega^* - \sum_{i=1}^{\frac{k-1}{2}}(4\sigma)^i\delta_1^{(2i)}(-x)\bigg).
\end{aligned}$$

Denote

$$r_1^*(x) = \eta^*\bigg(-\sum_{i=1}^{\frac{k-1}{2}}(4\sigma)^i\delta_1^{(2i-1)}(x)+\eta\bigg) + 4\sigma\omega^*\bigg(\omega^* - \sum_{i=1}^{\frac{k-1}{2}}(4\sigma)^i\delta_1^{(2i)}(-x)\bigg).$$

Next we prove $r_1^* > 0$. Notice that $\delta_1(1) > 0$, hence we have

$$\begin{aligned}
\omega^* - \sum_{i=1}^{\frac{k-1}{2}}(4\sigma)^i\delta_1^{(2i)}(-x) &> \sum_{i=1}^{\frac{k-1}{2}}(4\sigma)^i\sum_{l=2i}^{k}\frac{2l+1}{2}\bigg(p_l^{(2i)}(1) - p_l^{(2i)}(-x)\bigg) \\
&\geq \sum_{i=1}^{\frac{k-1}{2}}(4\sigma)^i\sum_{l=k-1}^{k}\frac{2l+1}{2}\bigg(p_l^{(2i)}(1) - p_l^{(2i)}(-x)\bigg) \\
&= r_1^{**}(x)/(4\sigma)\omega^*,
\end{aligned}$$

where

$$r_1^{**}(x)=(4\sigma)\omega^*\sum_{i=1}^{\frac{k-1}{2}}(4\sigma)^i\sum_{l=k-1}^{k}\frac{2l+1}{2}\bigg(p_l^{(2i)}(1)-p_l^{(2i)}(-x)\bigg).$$

Then we have

$$\begin{aligned}
r_1^*(x)&>\eta^*\bigg(-\sum_{i=1}^{\frac{k-1}{2}}(4\sigma)^i\delta_1^{(2i-1)}(x)+\eta\bigg)+r_1^{**}(x)\\
&=\sum_{i=1}^{\frac{k-1}{2}}(4\sigma)^i\sum_{l=0}^{k}\frac{2l+1}{2}\bigg(\delta_1^{(2i-1)}(-1)-\delta_1^{(2i-1)}(x)\bigg)\eta^*+r_1^{**}(x)\\
&=\sum_{i=1}^{\frac{k-1}{2}}(4\sigma)^i\bigg(\frac{1}{2}\sum_{l=k-1}^{k}\bigg(p_l^{(2i)}(-1)-p_l^{(2i)}(x)\bigg)+\frac{2k+1}{2}\bigg(p_k^{(2i-1)}(-1)-p_k^{(2i-1)}(x)\bigg)\bigg)\eta^*+r_1^{**}(x)\\
&\geq\sum_{i=1}^{\frac{k-1}{2}}(4\sigma)^i\bigg(\frac{1}{2}\sum_{l=k-1}^{k}\bigg(p_l^{(2i)}(-1)-p_l^{(2i)}(x)\bigg)+\frac{2k+1}{2}\bigg(p_k^{(2i-1)}(-1)-p_k^{(2i-1)}(x)\bigg)\bigg)\eta^*\\
&\quad+\sum_{i=1}^{\frac{k-1}{2}}(4\sigma)^i\sum_{l=k-1}^{k}\frac{2l+1}{2}\bigg(p_l^{(2i)}(1)-p_l^{(2i)}(-x)\bigg)4\sigma\omega^*.
\end{aligned}\tag{B.31}$$

By

$$\begin{aligned}
&p_{k-1}^{(2i)}(-x)=p_{k-1}^{(2i)}(x),\quad p_{k-1}^{(2i)}(-1)=p_{k-1}^{(2i)}(1),\\
&p_k^{(2i)}(-x)=-p_k^{(2i)}(x),\quad p_k^{(2i)}(-1)=-p_k^{(2i)}(1),\ p_k^{(2i-1)}(-1)=p_k^{(2i-1)}(1),
\end{aligned}$$

we have

$$\begin{aligned}
r_1^*(x)&\geq\bigg(\frac{1}{2}\eta^*+\frac{2k-1}{2}4\sigma\omega^*\bigg)\bigg(p_{k-1}^{(2i)}(1)-p_{k-1}^{(2i)}(x)\bigg)+\bigg(\frac{2k+1}{2}4\sigma\omega^*-\frac{1}{2}\eta^*\bigg)\bigg(p_k^{(2i)}(1)+p_k^{(2i)}(x)\bigg)\\
&\quad+\frac{2k+1}{2}\bigg(p_k^{(2i-1)}(1)-p_k^{(2i-1)}(x)\bigg)\eta^*.
\end{aligned}$$

Verifying

$$\frac{2k+1}{2}4\sigma\omega^*-\frac{1}{2}\eta^*>0\tag{B.32}$$

is enough. We have

$$\begin{aligned}
\eta^*&=\sum_{i=0}^{\frac{k-1}{2}}(4\sigma)^{i+1}\delta_1^{(2i+1)}(-1)=\sum_{i=0}^{\frac{k-1}{2}}\frac{(4\sigma)^{i+1}}{2}\sum_{l=k}^{k+1}p_l^{(2i+2)}(-1)=\sum_{i=0}^{\frac{k-1}{2}}\frac{(4\sigma)^{i+1}}{2}\bigg(p_{k+1}^{(2i+2)}(-1)+p_k^{(2i+2)}(-1)\bigg)\\
&=\sum_{i=0}^{\frac{k-1}{2}}\frac{(4\sigma)^{i+1}}{2}\bigg(\frac{1}{2^{(2i+2)}(2i+2)!}\frac{(k+1+2i+2)!}{(k-2i-1)!}-\frac{1}{2^{2i+2}(2i+2)!}\frac{(k+2i+2)!}{(k-2i-2)!}\bigg)
\end{aligned}$$

$$= \sum_{i=0}^{\frac{k-1}{2}} \frac{(4\sigma)^{i+1}}{2} \left( \frac{1}{2^{(2i+1)}(2i+1)!} \frac{(k+2i+2)!}{(k-2i-1)!} \right) = \sum_{i=0}^{\frac{k-1}{2}} \frac{(4\sigma)^{i+1}}{2} p_{k+1}^{(2i+1)}(1)(k-2i),$$

hence

$$(\frac{2k+1}{2})4\sigma\omega^* = \sum_{i=0}^{\frac{k-1}{2}} \frac{(4\sigma)^{i+1}}{2} \left( p_{k+1}^{(2i+1)}(1) + p_k^{(2i+1)}(1) \right) (\frac{2k+1}{2}) > \frac{\eta^*}{2}.$$

Now we have

$$r_1^*(x) > 0, \ r_1(x) > r_1^*(x) > 0, \ \forall x \in [-1,1]. \tag{B.33}$$

□

**Theorem B.2.** *Given LGL integral points set on reference cell* $\{x_\varrho\}_{\varrho=0}^k$ *where* $x_0 = -1$ *and* $x_k = 1$ *specially. Then* $r_2$ *and* $r_3$ *are positive when* $k \geq 3$ *and* $\sigma$ *satisfies the following conditions*

$$\sigma \geq \frac{1}{4} + \max_{\varrho=\{1,\cdots,k-1\}} \left( \frac{1}{1-|x_\varrho|} \left( \frac{4}{9} + \frac{2k+2}{2k+1} \right) \right). \tag{B.34}$$

**Proof:** To prove $r_2$ is positive, we view $r_2(\sigma, x)$ as a polynomial of $\sigma$ for fixed $x$. Notice that we only need to prove that $r_2 > 0$ holds on some fixed specific integral points as long as $\sigma$ satisfies some CFL conditions. Firstly, we prove that the leading term of $r_2$ is positive. Notice that for any fixed $x \in (-1,1)$, we have

$$\begin{aligned}
&r_2(\sigma,x) = \eta^* \left( \sum_{i_1=0}^{\frac{k-3}{2}} (4\sigma)^{i_1+1} \delta^{(2i_1+1)}(x) - \xi \right) + 4\sigma\omega^* \left( \sum_{i_1=1}^{\frac{k-1}{2}} (4\sigma)^{i_1} \left( \delta_1^{(2i_1)}(-x) - \delta_1^{(2i_1)}(-1) \right) \right) \\
&- \eta^* - 4\sigma\omega^* \delta_1(-1) \\
=&\sum_{i_1=1}^{\frac{k-1}{2}} \sum_{i_2=0}^{\frac{k-1}{2}} (4\sigma)^{i_1+i_2+1} \left( \left( \delta_1^{(2i_1-1)}(x) - \delta_1^{(2i_1-1)}(1) \right) \delta_1^{(2i_2+1)}(-1) \right. \\
&\left. + \left( \delta_1^{(2i_1)}(-x) - \delta_1^{(2i_1)}(-1) \right) \delta_1^{(2i_2)}(1) \right) - \eta^* - 4\sigma\omega^* \delta_1(-1) \\
=&\sum_{i_1=1}^{\frac{k-1}{2}} \sum_{i_2=0}^{\frac{k-1}{2}} (4\sigma)^{i_1+i_2+1} \left( \delta_1^{(2i_1)}(x^*) \delta_1^{(2i_2+1)}(-1) - \delta_1^{(2i_1+1)}(x^{**}) \delta_1^{(2i_2)}(1) \right) (x-1) - \eta^* - 4\sigma\omega^* \delta_1(-1),
\end{aligned}$$

by the mean value theorem, where $x^* \in (x,1)$ and $x^{**} \in (-1,-x)$. With Lemma A.3, we have the form of the leading term of $r_2$ as below:

$$\begin{aligned}
&(4\sigma)^k \left( \left( \delta_1^{(k-2)}(x) - \delta_1^{(k-2)}(1) \right) \delta_1^{(k)}(-1) + \delta_1^{(k-1)}(1) \left( \delta_1^{(k-1)}(-x) - \delta_1^{(k-1)}(-1) \right) \right) \\
=&(4\sigma)^k \left( \frac{1}{2} p_k^{(k-1)}(x) - \frac{1}{2} p_k^{(k-1)}(1) + \frac{2k+1}{2} p_k^{(k-2)}(x) - \frac{2k+1}{2} p_k^{(k-2)}(1) \right) p_k^{(k)} \frac{2k+1}{2} \\
&+ (4\sigma)^k (k+1) p_k^{(k)} \left( \frac{2k+1}{2} p_k^{(k-1)}(-x) - \frac{2k+1}{2} p_k^{(k-1)}(-1) \right) = (4\sigma)^k (\frac{2k+1}{2})^2 \frac{1}{2} (p_k^{(k)})^2 (1-x)^2.
\end{aligned}$$

Furthermore, we get

(B.35)

$$r_2(\sigma,x)=\Bigg(\sum_{i_1=1}^{\frac{k-3}{2}}\sum_{i_2=0}^{\frac{k-3}{2}}+\sum_{i_1\equiv\frac{k-1}{2},i_2=0}^{i_2=\frac{k-3}{2}}+\sum_{i_2\equiv\frac{k-1}{2},i_1=1}^{i_1=\frac{k-3}{2}}\Bigg)(4\sigma)^{i_1+i_2+1}\Bigg(\delta_1^{(2i_1)}(x^*)\delta_1^{(2i_2+1)}(-1)$$
$$-\delta_1^{(2i_1+1)}(x^{**})\delta_1^{(2i_2)}(1)\Bigg)(x-1)+(4\sigma)^k(\frac{2k+1}{2})^2\frac{1}{2}(p_k^{(k)})^2(1-x)^2-\eta^*-4\sigma\omega^*\delta_1(-1).$$

Note that

$$\delta_1^{(i)}(x)=\frac{1}{2}[p_k^{(i+1)}(x)+p_{k+1}^{(i+1)}(x)],\ p_{k+1}^{(k+1)}=(2k+1)p_k^{(k)},\ p_k^{(k)}\geq p_k^{(j)}(1)(k-j)!,$$

As a result, we have

$$\sum_{i_2=0}^{\frac{k-3}{2}}(4\sigma)^{i_2}\left|\delta_1^{(2i_2+1)}(x)\right|=\sum_{i_2=0}^{\frac{k-3}{2}}(4\sigma)^{i_2}\frac{1}{2}\left|[p_k^{(2i_2+1)}(x)+p_{k+1}^{(2i_2+1)}(x)]\right|$$
$$\leq\sum_{i_2=0}^{\frac{k-3}{2}}(4\sigma)^{i_2}p_{k+1}^{(2i_2+2)}(1)\leq\sum_{i_2=0}^{\frac{k-3}{2}}(4\sigma)^{i_2}p_{k+1}^{(k-1)}(1)$$
$$\leq\sum_{i_2=0}^{\frac{k-3}{2}}(4\sigma)^{i_2}\frac{1}{2}p_{k+1}^{(k+1)}=\frac{2k+1}{2}p_k^{(k)}\frac{(4\sigma)^{\frac{k-1}{2}}-1}{4\sigma-1}.$$

Similarly, we have

$$\sum_{i_1=1}^{\frac{k-3}{2}}(4\sigma)^{i_1}\left|\delta_1^{2i_1}(x)\right|\leq\frac{2k+1}{6}p_k^{(k)}\frac{(4\sigma)^{\frac{k-1}{2}}-1}{4\sigma-1},$$

and

$$(4\sigma)^{\frac{k-1}{2}}\left|\delta_1^{(k-1)}(x)\right|\leq(4\sigma)^{\frac{k-1}{2}}\frac{1}{2}[p_k^{(k)}+p_{k+1}^{(k)}(1)]\leq(k+1)p_k^{(k)}(4\sigma)^{\frac{k-1}{2}},$$
$$(4\sigma)^{\frac{k-1}{2}}\left|\delta_1^{(k)}(x)\right|=(4\sigma)^{\frac{k-1}{2}}\frac{2k+1}{2}p_k^{(k)}.$$

Then for any fixed $x_l\in\{x_\varrho\}_{\varrho=0}^k$, we get

$$r_2(\sigma,x_l)\geq(2k+1)^2(p_k^{(k)})^2(1-x_l)(4\sigma)^k\left(\frac{1}{8}(4\sigma-1)(1-x_l)-\frac{1}{6(4\sigma-1)}-\frac{1}{6}-\frac{k+1}{2k+1}\right)$$
$$-\eta^*-4\sigma\omega^*\delta_1(-1)\geq-\eta^*-4\sigma\omega^*\delta_1(-1),$$

as long as $\sigma$ satisfies the CFL condition defined in the theorem. Then we need to show that

$$-\eta^*-4\sigma\omega^*\delta_1(-1)>0.$$

Notice that

$$-\frac{2l+1}{2}p_l^{(2i_1+1)}(1)+\frac{2l+3}{2}p_{l+1}^{(2i_1+1)}(1)>0,$$

and

$$-\delta_1(-1) = \sum_{l=0}^{k} \frac{2l+1}{2}(-1)^{l+1} = \frac{k+1}{2},$$

therefore we have

$$\begin{aligned}
-\eta^* - 4\sigma\omega^*\delta_1(-1) &= \sum_{i_1=0}^{\frac{k-1}{2}} (4\sigma)^{i_1+1}\left(\frac{k+1}{2}\sum_{l=2i_1}^{k}\frac{2l+1}{2}p_l^{(2i_1)}(1) - \sum_{l=2i_1+1}^{k}\frac{2l+1}{2}(-1)^{l+1}p_l^{(2i_1+1)}(1)\right) \\
&> \sum_{i_1=0}^{\frac{k-1}{2}} (4\sigma)^{i_1+1}\left(\frac{k+1}{2}\sum_{l=2i_1}^{k}\frac{2l+1}{2}p_l^{(2i_1)}(1) - \frac{2k+1}{2}p_k^{(2i_1+1)}(1)\right) \\
&> \sum_{i_1=0}^{\frac{k-3}{2}} (4\sigma)^{i_1+1}\left(-\frac{2k+1}{2}p_k^{(2i_1+1)}(1)\right) + (4\sigma)^{\frac{k+1}{2}}\frac{2k+1}{2}p_k^{(k)}(1)(\frac{k+1}{2}-1) \\
&\geq p_k^{(k)}\left((4\sigma)^{\frac{k+1}{2}} - \frac{(4\sigma)^{\frac{k+1}{2}}-1}{4\sigma-1}\right) > 0,
\end{aligned}$$

as long as $\sigma \geq \frac{1}{2}$ and $k \geq 3$.

As for $r_3$ we can follow the same way. Firstly, the following inequality

$$\begin{aligned}
r_3(\sigma, x) = &\sum_{i_1=1}^{\frac{k-1}{2}}\sum_{i_2=0}^{\frac{k-1}{2}} (4\sigma)^{i_1+i_2+1}\left(\delta_1^{(2i_1)}(x^{\#})\delta_1^{(2i_2+1)} - \delta_1^{(2i_1+1)}(x^{\#\#})\delta_1^{(2i_2)}(-1)\right)(x+1) \\
&- 4\sigma\gamma^*\delta_1(1) + \left(\sum_{i=1}^{\frac{k-1}{2}}(4\sigma)^i\delta_1^{(2i-1)}(x) - \eta\right),
\end{aligned}$$

holds for any fixed $x \in (-1, 1)$ where $x^{\#} \in (x, 1)$ and $x^{\#\#} \in (-1, -x)$. The leading term of $r_3$ is

$$(4\sigma)^k(\frac{2k+1}{2})^2\frac{1}{2}(p_k^{(k)})^2(1+x)^2$$

and the $x$-part is the same as $r_2$. Then for any fixed $x_l \in \{x_\varrho\}_{\varrho=0}^{k}$, we have

$$r_3(\sigma, x_l) \geq -4\sigma\gamma^*\delta_1(1) + \left(\sum_{i=1}^{\frac{k-1}{2}}(4\sigma)^i\delta_1^{(2i-1)}(x_l) - \eta\right).$$

Notice that

$$\sum_{i=1}^{\frac{k-1}{2}}(4\sigma)^i\delta_1^{(2i-1)}(x_l) > 0,$$

therefore we only need to prove

$$-4\sigma\gamma^*\delta_1(1) - \eta \geq 0.$$

Note that $\delta_1^{(2i)}(-1) < 0$ and $\delta_1(1) = \frac{(k+1)^2}{2}$ and

$$|\delta_1^{(2i-1)}(-1)| \leq (k+1)p_k^{(k)},\ -\delta_1^{(k-1)}(-1) = kp_k^{(k)},$$

therefore it is sufficient to prove

$$k\frac{(k+1)^2}{2}(4\sigma)^{\frac{k+1}{2}} > (k+1)\sum_{i=1}^{\frac{k-1}{2}}(4\sigma)^i.$$

The above inequality is true clearly, then we prove that $r_3 > 0$.

B.3.1. *The* $\mathbf{f}^+$ *case.* Recall our aim of representing the integral using point values, we then prove when $x = -1$,

$$w_0 r_1(\sigma,-1) - 2\sigma(\gamma^* + \omega^*) > 0, \tag{B.36}$$

$$w_0 r_2(\sigma,-1) + 2\sigma\omega^* > 0, \tag{B.37}$$

$$w_0 r_3(\sigma,-1) + 2\sigma\gamma^* > 0. \tag{B.38}$$

For the first inequality (B.36), we have

$$w_0 r_1(\sigma,-1) - 2\sigma(\gamma^* + \omega^*) > \frac{w_0}{2}4\sigma\omega^*\delta_1(1) - 2\sigma\omega^* = (\frac{w_0}{2}(k+1)^2 - 1)2\sigma\omega^*.$$

Notice that if we adopt the LGL points, then we get the coefficient $w_0 = \frac{2}{k(k+1)}$ (see [25]) and

$$w_0 r_1(\sigma,-1) - 2\sigma(\gamma^* + \omega^*) > 0$$

holds directly. For the second inequality (B.37), recall that

$$\begin{aligned} 2r_2(\sigma,-1) = &\sum_{i_1=1}^{\frac{k-1}{2}}\sum_{i_2=0}^{\frac{k-1}{2}}(4\sigma)^{i_1+i_2+1}\Big(\delta_1^{(2i_1-1)}(-1)\delta_1^{(2i_2+1)}(-1) - \delta_1^{(2i_1-1)}(1)\delta_1^{(2i_2+1)}(-1) \\ &-\delta_1^{(2i_1)}(-1)\delta_1^{(2i_2)}(1) + \delta_1^{(2i_1)}(1)\delta_1^{(2i_2)}(1)\Big) - \delta_1(-1)\sum_{i_2=0}^{\frac{k-1}{2}}(4\sigma)^{i_2+1}\delta_1^{(2i_2)}(1) \\ &+\delta_1(1)\sum_{i_2=0}^{\frac{k-1}{2}}(4\sigma)^{i_2+1}\delta_1^{(2i_2)}(1) - \eta^*. \end{aligned}$$

Notice that

$$\delta_1(1) = \sum_{l=0}^{k}\frac{2l+1}{2} = \frac{(k+1)^2}{2},$$

and, by the third property of Lemma A.1, we have

$$\frac{(k+1)^2}{2}p_l^{(2i_2)}(1) = p_l^{(2i_2+1)}(1)\frac{(k+1)^2(2i_2+1)}{(l+2i_2+1)(l-2i_2)} \geq (\frac{k+1}{k+\frac{1}{2}})^2 p_l^{(2i_2+1)}(1) > p_l^{(2i_2+1)}(-1).$$

Hence we have

$$\delta_1(1)\sum_{i_2=0}^{\frac{k-1}{2}}(4\sigma)^{i_2+1}\delta_1^{(2i_2)}(1) > \eta^*.$$

Also, by Lemma A.3 and the first property of Lemma A.1, we have the following equations:

$$\delta_1^{(k-2)}(-1) = -\frac{1}{2}p_k^{(k-1)}(1) + \frac{1}{2}p_{k+1}^{(k-1)}(1),\ \delta_1^{(k-2)}(1) = \frac{1}{2}p_k^{(k-1)}(1) + \frac{1}{2}p_{k+1}^{(k-1)}(1),$$
$$\delta_1^{(k)}(-1) = \frac{2k+1}{2}p_k^{(k)},\ \delta_1^{(k-1)}(-1) = \frac{1}{2}p_k^{(k)} - \frac{1}{2}p_{k+1}^{(k+1)},\ \delta_1^{(k-1)}(1) = \frac{1}{2}p_k^{(k)} + \frac{1}{2}p_{k+1}^{(k+1)}.$$

Therefore the coefficient of $(4\sigma)^k$ of $2r_2(\sigma,-1)$ has the following representation:

$$\left(\delta_1^{(k-2)}(-1) - \delta_1^{(k-2)}(1)\right)\delta_1^{(k)}(-1) - \left(\delta_1^{(k-1)}(-1) + \delta_1^{(k-1)}(1)\right)\delta_1^{(k-1)}(1) = \frac{1}{2}\left(p_{k+1}^{(k+1)}\right)^2.$$

Notice that $\delta_1(-1) > 0$. Hence we have

$$2r_2(\sigma,-1) > \sum_{i_1=1}^{\frac{k-3}{2}}\sum_{i_2=0}^{\frac{k-3}{2}}(4\sigma)^{i_1+i_2+1}\left(\delta_1^{(2i_1-1)}(-1) - \delta_1^{(2i_1-1)}(1)\right)\delta_1^{(2i_2+1)}(-1) + \frac{1}{2}(4\sigma)^k\left(p_{k+1}^{(k+1)}\right)^2.$$

Furthermore, notice that

$$\delta_1^{(2i_1-1)}(-1) - \delta_1^{(2i_1-1)}(1) = -p_k^{(2i_1)}(1) \le \frac{1}{6}p_{k+1}^{(k+1)},\ i_1 = 0,\cdots,\frac{k-3}{2},$$

and

$$\delta_1^{(2i_1+1)}(-1) = \frac{1}{2}\left(p_k^{(2i_1+2)}(-1) + p_{k+1}^{(2i_1+2)}(-1)\right) \le p_{k+1}^{(k+1)},\ i_1 = 0,\cdots,\frac{k-3}{2}.$$

Therefore we have

$$2r_2(\sigma,-1) > (4\sigma)^k\left(p_{k+1}^{(k+1)}\right)^2\left(\frac{1}{2} - \frac{1}{6}\frac{1}{(4\sigma-1)^2}\right) > 0.$$

Now we have the second inequality. For the third inequality (B.38), recall that

$$r_3(\sigma,-1) = -2\sigma\gamma^*(\delta_1(1)w_0 - 1) > 0.$$

Hence we get the coefficient matrix of the vector $\mathbf{f}^+$ is positive.

B.3.2. *The* $\mathbf{f}^-$ *case.* It is enough to only prove $r_3(\sigma,1) > 0$ here. Recall that

$$r_3(\sigma,1) = \frac{1}{2}\left(\sum_{i=1}^{\frac{k-1}{2}}(4\sigma)^i\delta_1^{(2i-1)}(1) - \eta\right)(1+\xi^*) + 2\sigma\gamma^*\left(\sum_{i=1}^{\frac{k-1}{2}}(4\sigma)^i\delta_1^{(2i)}(-1) - \omega^*\right).$$

Because of

$$\sum_{i=1}^{\frac{k-1}{2}}(4\sigma)^i\delta_1^{(2i-1)}(1) - \eta = \sum_{i=1}^{\frac{k-1}{2}}(4\sigma)^i\delta_1^{(2i-1)}(1) - \sum_{i=1}^{\frac{k-1}{2}}(4\sigma)^i\delta_1^{(2i-1)}(-1) > 0,\ \xi^* > 0,$$

and

$$\sum_{i=1}^{\frac{k-1}{2}}(4\sigma)^i\delta_1^{(2i)}(-1) - \omega^* = \sum_{i=1}^{\frac{k-1}{2}}(4\sigma)^i\delta_1^{(2i)}(-1) - \sum_{i=0}^{\frac{k-1}{2}}(4\sigma)^i\delta_1^{(2i)}(1) < 0,\ \gamma^* < 0,$$

we have $r_3(\sigma,1) > 0$. Hence we get the coefficient matrix of the vector $\mathbf{f}^-$ is positive. □

B.4. **Proof of Theorem 2.7.**

**Proof:** In our case, we take $\zeta_1 = 1,\ \zeta_2 = -\frac{3\sigma-144\sigma^2+3600\sigma^3}{1+162\sigma+3312\sigma^2+7200\sigma^3},\ \zeta_3 = 0$ in Lemma 2.1. It is easy to check the condition (2.35) holds. Recall that $T^i(z)$ and $U^i(z)$ are the first and the second classes of Chebyshev polynomials and we have

$$\begin{aligned} T^i(z) &= \frac{1}{2}\left(\left(z+\sqrt{z^2-1}\right)^i + \left(z-\sqrt{z^2-1}\right)^i\right), \\ U^i(z) &= \frac{1}{2\sqrt{z^2-1}}\left(\left(z+\sqrt{z^2-1}\right)^{i+1} - \left(z-\sqrt{z^2-1}\right)^{i+1}\right). \end{aligned} \tag{B.39}$$

Then by Lemma 2.1 we have

$$(1+162\sigma+3312\sigma^2+7200\sigma^3)(\mathbf{T}_1-60\sigma\mathbf{T}_4)^{-1} = \sum_{s=0}^{N-1} g_s\mathbf{G}^s, \tag{B.40}$$

where

$$\begin{aligned} g_s(z) = &\frac{1}{(-\zeta_2)\cdot 2[T^N(z)-1]\cdot 2\sqrt{z^2-1}} \\ &\cdot\left(\left(z+\sqrt{z^2-1}\right)^s - \left(z+\sqrt{z^2-1}\right)^{-s} + \left(z+\sqrt{z^2-1}\right)^{N-s} - \left(z+\sqrt{z^2-1}\right)^{-(N-s)}\right), \end{aligned} \tag{B.41}$$

for each $s = 0, 1, \cdots, N-1$ with $z = \frac{\zeta_3-\zeta_1}{2(\zeta_2-\zeta_3)}$. Notice that we take a modification of index and $s = i_1 - 1$ in Lemma 2.1 indeed. Then (2.31) can be reformulated as:

$$\bar{\mathbf{u}} = (\mathbf{T}_1-60\sigma\mathbf{T}_4)^{-1}\mathbf{Q}_1\mathbf{f}^+ + (\mathbf{T}_1-60\sigma\mathbf{T}_4)^{-1}\mathbf{Q}_2\mathbf{f}^0 + (\mathbf{T}_1-60\sigma\mathbf{T}_4)^{-1}\mathbf{Q}_3\mathbf{f}^-. \tag{B.42}$$

Now we are aiming to prove that:

$$(\mathbf{T}_1-60\sigma\mathbf{T}_4)^{-1}\mathbf{Q}_i > 0, \quad i = 1, 2, 3.$$

To compute $\mathbf{Q}_i$, we calculate $\mathbf{T}_3^{-1}(-3\mathbf{T}_4\mathbf{B}^\top + \mathbf{T}_1\mathbf{A})$ first. Take $r = \frac{24\sigma}{1+36\sigma} < 1$, and we know the following equation by von Neumann expansion:

$$\mathbf{T}_3^{-1} = \sum_{i'=0}^{N} a_{i'}\mathbf{G}^{N-i'}, \tag{B.43}$$

where

$$a_{i'} = \sum_{s=0}^{\infty}(-r)^{Ns+i'} = \frac{(-r)^{i'}}{1-(-r)^N}. \tag{B.44}$$

Then we calculate $-3\mathbf{T}_4\mathbf{B}^\top + \mathbf{T}_1\mathbf{A}$:

$$-3\mathbf{T}_4\mathbf{B}^\top = -(117\sigma+900\sigma^2)\mathbf{I}+(27\sigma+180\sigma^2)\mathbf{G}+(117\sigma+1260\sigma^2)\mathbf{G}^{N-1}-(27\sigma+540\sigma^2)\mathbf{G}^{N-2},$$

$$\mathbf{T}_1\mathbf{A} = (1+165\sigma+1908\sigma^2)\mathbf{I}+(-3\sigma+684\sigma^2)\mathbf{G}-(2268\sigma^2+165\sigma+1)\mathbf{G}^{N-1}-(-3\sigma+324\sigma^2)\mathbf{G}^{N-2}.$$

Add the above two equations, we have

$$-3\mathbf{T}_4\mathbf{B}^\top + \mathbf{T}_1\mathbf{A} = (1008\sigma^2+48\sigma+1)\left(\mathbf{I}-\mathbf{G}^{N-1}\right) + (864\sigma^2+24\sigma)\left(\mathbf{G}-\mathbf{G}^{N-2}\right). \tag{B.45}$$

Then we have

(B.46)

$$\begin{aligned}
&\mathbf{T}_3^{-1}(-3\mathbf{T}_4\mathbf{B}^\top + \mathbf{T}_1\mathbf{A}) \\
&= \frac{1}{1+36\sigma}\sum_{i'=0}^{N-1} a_{i'}\mathbf{G}^{N-i'}\Big((1008\sigma^2+48\sigma+1)\Big(\mathbf{I}-\mathbf{G}^{N-1}\Big) + (864\sigma^2+24\sigma)\Big(\mathbf{G}-\mathbf{G}^{N-2}\Big)\Big) \\
&= \sum_{i'=0}^{N-1}\left(\frac{1008\sigma^2+48\sigma+1}{1+36\sigma}(a_{i'}-a_{i'-1}) + 24\sigma(a_{i'+1}-a_{i'-2})\right)\mathbf{G}^{N-i'} = \sum_{i'=0}^{N-1} a_{i'}^*\mathbf{G}^{N-i'},
\end{aligned}$$

where

$$a_{i'}^* = \sum_{i'=0}^{N-1}\left(\frac{1008\sigma^2+48\sigma+1}{1+36\sigma}(a_{i'}-a_{i'-1}) + 24\sigma(a_{i'+1}-a_{i'-2})\right).$$

Next we will show that the $a_{i'}^*$ has an extremely simple form:

$$a_{i'}^* = \begin{cases} 1+12\sigma, & i'=0, \\ -(1+36\sigma), & i'=1, \\ 0, & 2\le i'\le N-2, \\ 24\sigma, & i'=N-1. \end{cases} \tag{B.47}$$

To prove the above equation, we put $a_{i'}$ into $a_{i'}^*$, when $2\le i'\le N-2$ we have

(B.48)

$$\begin{aligned}
a_{i'}^* &= \frac{1008\sigma^2+48\sigma+1}{1+36\sigma}(a_{i'}-a_{i'-1}) + 24\sigma(a_{i'+1}-a_{i'-2}) \\
&= (-r)^{i'}\left(\frac{1008\sigma^2+48\sigma+1}{1+36\sigma}\times\frac{1-(\frac{1}{-r})}{1-(-r)^N} + 24\sigma\times\frac{(-r)-\frac{1}{r^2}}{1-(-r)^N}\right) \\
&= \frac{(-r)^{i'}}{(1-(-r)^N)24\sigma(1+36\sigma)}\Big((1008\sigma^2+48\sigma+1)(1+60\sigma) - ((1+36\sigma)^3+(24\sigma)^3)\Big) = 0.
\end{aligned}$$

In the case of $i'=1$ or $i'=N-1$, we have

(B.49)

$$\begin{aligned}
a_1^* &= \frac{1008\sigma^2+48\sigma+1}{1+36\sigma}(a_1-a_0) + 24\sigma(a_2-a_{N-1}) \\
&= \underbrace{\frac{1008\sigma^2+48\sigma+1}{1+36\sigma}(a_1-a_0) + 24\sigma(a_2-\frac{(-r)^{-1}}{1-(-r)^N})}_{0} + 24\sigma\left(\frac{(-r)^{-1}}{1-(-r)^N}\right) - 24\sigma a_{N-1} \\
&= -(1+36\sigma),
\end{aligned}$$

and

(B.50)
$$\begin{aligned} a_{N-1}^* &= \frac{1008\sigma^2+48\sigma+1}{1+36\sigma}(a_{N-1}-a_{N-2})+24\sigma(a_0-a_{N-3}) \\ &= \underbrace{\frac{1008\sigma^2+48\sigma+1}{1+36\sigma}(a_{N-1}-a_{N-2})+24\sigma(\frac{(-r)^N}{1-(-r)^N}-a_{N-3})}_{0} -24\sigma\left(\frac{(-r)^N}{1-(-r)^N}\right)+24\sigma a_0 \\ &= 24\sigma. \end{aligned}$$

When $i'=0$, we have

$$\begin{aligned} a_0^* &= \frac{1008\sigma^2+48\sigma+1}{1+36\sigma}(a_0-a_{N-1})+24\sigma(a_1-a_{N-2}), \\ a_1^* &= \frac{1008\sigma^2+48\sigma+1}{1+36\sigma}(a_1-a_0)+24\sigma(a_2-a_{N-1}), \end{aligned}$$

(B.51)
$$\vdots$$

$$\begin{aligned} a_{N-2}^* &= \frac{1008\sigma^2+48\sigma+1}{1+36\sigma}(a_{N-2}-a_{N-3})+24\sigma(a_{N-1}-a_{N-4}), \\ a_{N-1}^* &= \frac{1008\sigma^2+48\sigma+1}{1+36\sigma}(a_{N-1}-a_{N-2})+24\sigma(a_0-a_{N-3}). \end{aligned}$$

Add all the above equations, we have

$$\sum_{i'=0}^{N-1} a_{i'}^* = 0.$$

As a result, we get

(B.52)
$$a_0^* = -(a_1^*+a_{N-1}^*) = 1+12\sigma.$$

Now we finish the proof. Introduce a new variable as below for easiness:

(B.53)
$$\mathbf{T}_5 = \mathbf{T}_3^{-1}(-3\mathbf{T}_4\mathbf{B}^\top+\mathbf{T}_1\mathbf{A}) = (1+12\sigma)\mathbf{I}+24\sigma\mathbf{G}-(1+36\sigma)\mathbf{G}^{N-1}.$$

Then we get

$$\begin{aligned} \mathbf{Q}_1 &= \frac{1}{6}\mathbf{T}_1+\mathbf{T}_4+4\sigma\mathbf{T}_5 \\ &= \underbrace{\left(360\sigma^2+19\sigma+\frac{1}{6}\right)}_{q_{11}}\mathbf{I}+\underbrace{\left(270\sigma^2+\frac{17}{2}\sigma\right)}_{q_{12}}\mathbf{G}+\underbrace{\left(-30\sigma^2-\frac{3\sigma}{2}\right)}_{q_{13}}\mathbf{G}^{N-1}, \\ \mathbf{Q}_2 &= \frac{2}{3}\mathbf{T}_1-4\sigma\mathbf{T}_5 \\ &= \underbrace{\left(1680\sigma^2+104\sigma+\frac{2}{3}\right)}_{q_{21}}\mathbf{I}+\underbrace{\left(360\sigma^2-2\sigma\right)}_{q_{22}}\mathbf{G}+\underbrace{\left(360\sigma^2+2\sigma\right)}_{q_{23}}\mathbf{G}^{N-1}, \\ \mathbf{Q}_3 &= \frac{1}{6}\mathbf{T}_1-6\sigma\mathbf{T}_5 \end{aligned} \tag{B.54}$$

$$= \underbrace{\left(360\sigma^2 + 21\sigma + \frac{1}{6}\right)}_{q_{31}}\mathbf{I} + \underbrace{\left(-30\sigma^2 - \frac{\sigma}{2}\right)}_{q_{32}}\mathbf{G} + \underbrace{\left(270\sigma^2 + \frac{11}{2}\sigma\right)}_{q_{33}}\mathbf{G}^{N-1}.$$

We already know that if $\sigma > 0.056$, then we have $\mathbf{Q}_2 > 0$ and $\mathbf{T}_1 - 60\sigma\mathbf{T}_4$ is an $M$-matrix. It is also easy to check all row sums of matrices $\mathbf{M}$ and $\sum_{i=1}^{3}\mathbf{Q}_i$ are equal and equal to $1 + 156\sigma + 3600\sigma^2$. Then we only need to derive the conditions which make:

$$(\mathbf{T}_1 - 60\sigma\mathbf{T}_4)^{-1}\mathbf{Q}_i > 0, \quad i = 1, 3. \tag{B.55}$$

**Remark**: To prove (B.55), we omit the positive term $1 + 162\sigma + 3312\sigma^2 + 7200\sigma^3$ for simplicity.

Notice that for $i = 1, 3$, we have

$$\begin{aligned}(\mathbf{T}_1 - 60\sigma\mathbf{T}_4)^{-1}\mathbf{Q}_i &= \sum_{s=0}^{N-1} g_s\mathbf{G}^s\left(q_{i1}\mathbf{I} + q_{i2}\mathbf{G} + q_{i3}\mathbf{G}^{N-1}\right)\\ &= \sum_{s=0}^{N-1}\left(g_s q_{i1} + g_{s-1}q_{i2} + g_{s+1}q_{i3}\right)\mathbf{G}^s.\end{aligned} \tag{B.56}$$

Now we want to get

$$g_s q_{i1} + g_{s-1}q_{i2} + g_{s+1}q_{i3} > 0, \quad i = 1, 3, \ s = 0, 1, \cdots, N-1, \tag{B.57}$$

where $g_N = g_0,\ g_{-1} = g_{N-1}$, and

$$\begin{aligned}q_{11} &= 360\sigma^2 + 19\sigma + \frac{1}{6},\ q_{12} = 270\sigma^2 + \frac{17}{2}\sigma,\ q_{13} = -30\sigma^2 - \frac{3\sigma}{2},\\ q_{31} &= 360\sigma^2 + 21\sigma + \frac{1}{6},\ q_{32} = -30\sigma^2 - \frac{\sigma}{2},\ q_{33} = 270\sigma^2 + \frac{11}{2}\sigma.\end{aligned}$$

Recall the definition of $g_s$:

$$\begin{aligned}g_s(z) = &\frac{1}{(-\zeta_2)\cdot 2[T^N(z) - 1]\cdot 2\sqrt{z^2-1}}\\ &\cdot\left(\left(z + \sqrt{z^2-1}\right)^s - \left(z + \sqrt{z^2-1}\right)^{-s} + \left(z + \sqrt{z^2-1}\right)^{N-s} - \left(z + \sqrt{z^2-1}\right)^{-(N-s)}\right),\end{aligned}$$

where $\zeta_2 = -\frac{3\sigma - 144\sigma^2 + 3600\sigma^3}{1 + 162\sigma + 3312\sigma^2 + 7200\sigma^3},\ z = -\frac{1}{2\zeta_2}$. It is clear that $-\zeta_2 \in (0, \frac{1}{2}),\ z \in (1, \infty)$ if $\sigma > 0$. Furthermore, we have

$$\frac{1}{(-\zeta_2)\cdot 2[T^n(z) - 1]\cdot 2\sqrt{z^2-1}} > 0.$$

Firstly we prove $(\mathbf{T}_1 - 60\sigma\mathbf{T}_4)^{-1}\mathbf{Q}_1$ is positive. Denote $\nu = z + \sqrt{z^2-1}$. Because $g_s$ is equal to $g_{N-s}$, we do not need to consider $s = 0$ and $s = N-1$ individually. For each $s = 0, \cdots, N-1$, we have

$$\begin{aligned}&g_s q_{11} + g_{s-1}q_{12} + g_{s+1}q_{13}\\ =&\nu^s\left(q_{11} + \frac{1}{\nu}q_{12} + \nu q_{13}\right) - \nu^{-s}\left(q_{11} + \nu q_{12} + \frac{1}{\nu}q_{13}\right)\\ &+ \nu^{N-s}\left(q_{11} + \nu q_{12} + \frac{1}{\nu}q_{13}\right) - \nu^{s-N}\left(q_{11} + \frac{1}{\nu}q_{12} + \nu q_{13}\right)\end{aligned}$$

$$=(\nu^s-\nu^{s-N})\left(q_{11}+\frac{1}{\nu}q_{12}+\nu q_{13}\right)+(\nu^{N-s}-\nu^{-s})\left(q_{11}+\nu q_{12}+\frac{1}{\nu}q_{13}\right).$$

Notice that $q_{13}<0$, and $|q_{13}|<|q_{11}|,\ \nu>1$. Therefore we focus on proving

$$q_{11}+\frac{1}{\nu}q_{12}+\nu q_{13}>0.$$

We have

$$q_{11}+\frac{1}{\nu}q_{12}+\nu q_{13}=\underbrace{\left(360\sigma^2+19\sigma+\frac{1}{6}\right)}_{\text{I}}+\underbrace{\left(\frac{270}{\nu}\sigma^2+\frac{17}{2\nu}\sigma\right)}_{\text{II}}+\underbrace{\left(-30\nu\sigma^2-\frac{3\nu}{2}\sigma\right)}_{\text{III}},$$

and the II part is positive. It is also easy to see that if $\nu<12$ then the sum of I and III is positive, and $\nu<12$ holds when $\sigma>0.162$ which means that:

$$\begin{aligned}\nu<12 &\iff z+\sqrt{z^2-1}<12\\ &\iff z<\frac{145}{24}\\ &\iff -\zeta_2>\frac{12}{145}\\ &\iff \frac{3\sigma-144\sigma^2+3600\sigma^3}{1+162\sigma+3312\sigma^2+7200\sigma^3}>\frac{12}{145}\\ &\Longleftarrow \sigma>0.162.\end{aligned}$$

Then we obtain the results in the case of $s=0,\cdots,N-1$ if $\sigma>0.162$. Next we will show that the $(\mathbf{T}_1-60\sigma\mathbf{T}_4)^{-1}\mathbf{Q}_3$ is positive.

As before, if $s=0,1,\cdots,N-1$, we have

(B.58)

$$\begin{aligned}g_sq_{31}+g_{s-1}q_{32}+g_{s+1}q_{33}=&\nu^s\left(q_{31}+\frac{1}{\nu}q_{32}+\nu q_{33}\right)-\nu^{-s}\left(q_{31}+\nu q_{32}+\frac{1}{\nu}q_{33}\right)\\ &+\nu^{N-s}\left(q_{31}+\nu q_{32}+\frac{1}{\nu}q_{33}\right)-\nu^{s-N}\left(q_{31}+\frac{1}{\nu}q_{32}+\nu q_{33}\right).\end{aligned}$$

Notice that $q_{32}<0$, and $|q_{32}|<|q_{31}|,\ \nu>1$, then we just need to prove

$$q_{31}+\nu q_{32}+\frac{1}{\nu}q_{33}>0. \tag{B.59}$$

We find that, if $\nu<12$, then we can get the conclusion in the same way. It is easy to check the row sums of $\mathbf{T}_1-60\sigma\mathbf{T}_4$ are equal to the row sums of $\mathbf{Q}_i$. Then in the same way as in Theorem (2.5), we get the condition of the MPP property of $P^2$ implicit LDG for the diffusion equation which is $\sigma>0.162$. □

B.5. **Definitions in Theorem 2.8.** The definitions of the terms in Theorem 2.8 are as below

$$\mathbf{A}_N=\mathbf{I}_N-\mathbf{G}_N^{-1},\ \mathbf{B}_N=2\mathbf{I}_N+4\mathbf{G}_N,$$
$$\mathbf{T}_{1,x}=6\mathbf{A}_{N_x}^\top+\mathbf{B}_{N_x}^\top\mathbf{B}_{N_x},\ \mathbf{T}_{1,y}=6\mathbf{A}_{N_y}^\top+\mathbf{B}_{N_y}^\top\mathbf{B}_{N_y},$$
$$\mathbf{T}_1=\mathbf{I}_{N_xN_y}+\sigma_x(\mathbf{I}_{N_y}\bigotimes\mathbf{T}_{1,x})+\sigma_y(\mathbf{T}_{1,y}\bigotimes\mathbf{I}_{N_x}),$$

$$\mathbf{T}_2^1 = 6\sigma_x(\mathbf{I}_{Ny} \bigotimes \mathbf{B}_{N_x}^\top),\ \mathbf{T}_2^2 = 6\sigma_y(\mathbf{B}_{N_y}^\top \bigotimes \mathbf{I}_{N_x}),$$
$$\mathbf{T}_3 = \mathbf{I}_{N_xN_y} + 6\sigma_x(\mathbf{I}_{N_y} \bigotimes \mathbf{A}_{N_x}) + 6\sigma_y(\mathbf{A}_{N_y} \bigotimes \mathbf{I}_{N_x}),$$
$$\mathbf{T}_4^1 = \sigma_x(\mathbf{I}_{N_y} \bigotimes \mathbf{A}_{N_x}\mathbf{B}_{N_x}),\ \mathbf{T}_4^2 = \sigma_y(\mathbf{A}_{N_y}\mathbf{B}_{N_y} \bigotimes \mathbf{I}_{N_x}),$$
$$\mathbf{T}_5 = \mathbf{I}_{N_y} \bigotimes \mathbf{I}_{N_x} + \sigma_x\mathbf{I}_{N_y} \bigotimes \mathbf{T}_{1,x} + 6\sigma_y\mathbf{A}_{N_y} \bigotimes \mathbf{I}_{N_x},$$
$$\mathbf{T}_6 = \mathbf{I}_{N_y} \bigotimes \mathbf{I}_{N_x} + \sigma_y\mathbf{T}_{1,y} \bigotimes \mathbf{I}_{N_x} + 6\sigma_x\mathbf{I}_{N_y} \bigotimes \mathbf{A}_{N_x},$$
$$\mathbf{P}_1 = \mathbf{T}_1\mathbf{T}_5 - \mathbf{T}_4^2\mathbf{T}_2^2,\ \mathbf{P}_2 = \mathbf{T}_1\mathbf{T}_6 - \mathbf{T}_4^1\mathbf{T}_2^1,$$
$$\mathbf{P}_3 = \mathbf{P}_2\mathbf{T}_1\mathbf{T}_2^1 + \mathbf{T}_4^2\mathbf{T}_2^1\mathbf{T}_1\mathbf{T}_2^2,\ \mathbf{P}_4 = \mathbf{P}_1\mathbf{T}_1\mathbf{T}_2^2 + \mathbf{T}_4^1\mathbf{T}_2^2\mathbf{T}_1\mathbf{T}_2^1,$$
$$\mathbf{P}_5 = \mathbf{P}_2\mathbf{T}_4^2 + \mathbf{T}_4^2\mathbf{T}_2^1\mathbf{T}_4^1,\ \mathbf{P}_6 = \mathbf{P}_1\mathbf{T}_4^1 + \mathbf{T}_4^1\mathbf{T}_2^2\mathbf{T}_4^2,$$
$$\mathbf{Q}_1 = \mathbf{T}_4^1\mathbf{P}_2\mathbf{T}_1 + \mathbf{T}_4^2\mathbf{T}_4^1\mathbf{T}_2^2\mathbf{T}_1,\ \mathbf{Q}_2 = \mathbf{T}_4^1\mathbf{T}_4^2\mathbf{T}_2^1\mathbf{T}_1 + \mathbf{T}_4^2\mathbf{P}_1\mathbf{T}_1,\ \mathbf{Q}_3 = \mathbf{T}_4^1\mathbf{P}_5 + \mathbf{T}_4^2\mathbf{P}_6,$$
$$\mathbf{Q} = \mathbf{P}_1\mathbf{P}_2 - \mathbf{T}_4^1\mathbf{T}_4^2\mathbf{T}_2^1\mathbf{T}_2^2,\ \ \mathbf{Q}_4 = \mathbf{T}_1^{-1}\mathbf{Q},\ \mathbf{F} = \mathbf{T}_1^{-1}\left(\mathbf{Q}\mathbf{T}_3 - \mathbf{T}_4^1\mathbf{P}_3 - \mathbf{T}_4^2\mathbf{P}_4\right),$$
$$\mathbf{W}_1 = \frac{1}{2}(\mathbf{Q}_1 + \mathbf{Q}_2) + \mathbf{Q}_3 + \frac{1}{4}\mathbf{Q}_4,\ \mathbf{W}_2 = \frac{1}{2}\mathbf{Q}_2 + \frac{1}{4}\mathbf{Q}_4,\ \mathbf{W}_3 = \frac{1}{2}\mathbf{Q}_1 + \frac{1}{4}\mathbf{Q}_4,\ \mathbf{W}_4 = \frac{1}{4}\mathbf{Q}_4.$$

B.6. **Proof of Theorem 2.8.** Denote that $\bar{\mathbf{u}} = (\bar{u}_{11}, \cdots, \bar{u}_{N_x1}, \bar{u}_{21}, \cdots, \bar{u}_{N_xN_y})$, with $\bar{u} = \frac{1}{|\Omega_{ij}|}\int_{\Omega_{ij}} u d\Omega$, and the elements of $\mathbf{f}_{x^+}/h_y$ are the corresponding line averages at the cell boundaries, i.e., $(\mathbf{f}_{x^+})_{(j-1)N_x+i} = \int_{y_{j-\frac{1}{2}}}^{y_{j+\frac{1}{2}}} f(x_{i-\frac{1}{2}}^+, y)dy$ and $(\mathbf{f}_{y^+})_{(j-1)N_x+i} = \int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} f(x, y_{j-\frac{1}{2}}^+)dx$. Similarly we define the vector $\bar{\mathbf{f}}$, $\bar{\mathbf{q}}^1$, $\bar{\mathbf{q}}^2$, $\mathbf{q}_{x^\pm}^1$, $\mathbf{q}_{y^\pm}^2$, $\mathbf{u}_{x^+}$, $\mathbf{u}_{y^+}$. Also we define $\mathbf{u}_\varrho$, $\mathbf{q}_\varrho^1$ and $\mathbf{q}_\varrho^2$ through the same way as $\mathbf{f}_\varrho$, for $\varrho = 1, 2, 3, 4$, which are defined in Theorem 2.9. We take uniform meshes in the $x$-axis and the $y$-axis separately. Take $v = 1$ in (2.39a), we have

(B.60)
$$\frac{h_xh_y}{\tau}[\bar{u}_{ij} - \bar{f}_{ij}] = \int_{y_{j-\frac{1}{2}}}^{y_{j+\frac{1}{2}}} q^1(x_{i+\frac{1}{2}}^-, y) - q^1(x_{i-\frac{1}{2}}^-, y)dy + \int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} q^2(x, y_{j+\frac{1}{2}}^-) - q^2(x, y_{j-\frac{1}{2}}^-)dx.$$

Take $w = \delta_{1,i}(x)$ in (2.39b) with Lemma A.1, we have

(B.61)
$$\int_{y_{j-\frac{1}{2}}}^{y_{j+\frac{1}{2}}} q^1(x_{i+\frac{1}{2}}^-, y)dy = -\frac{6h_y}{h_x}\bar{u}_{ij} + \frac{4}{h_x}\left(\int_{y_{j-\frac{1}{2}}}^{y_{j+\frac{1}{2}}} u(x_{i+\frac{1}{2}}^+, y)dy\right) + \frac{2}{h_x}\left(\int_{y_{j-\frac{1}{2}}}^{y_{j+\frac{1}{2}}} u(x_{i-\frac{1}{2}}^+, y)dy\right).$$

Similarly taking $\zeta = \delta_{1,j}(y)$, we have

(B.62)
$$\int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} q^2(x, y_{j+\frac{1}{2}}^-)dx = -\frac{6h_x}{h_y}\bar{u}_{ij} + \frac{4}{h_y}\left(\int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} u(x, y_{j+\frac{1}{2}}^+)dx\right) + \frac{2}{h_y}\left(\int_{x-\frac{1}{2}}^{x_{i+\frac{1}{2}}} u(x, y_{j-\frac{1}{2}}^+)dx\right).$$

To eliminate the boundary integral of $u$, we take $v = \delta_{-1,i}(x)$ and $v = \delta_{-1,j}(y)$ separately in (2.39a), then we have

(B.63)
$$\begin{aligned}&\frac{1}{\tau}\bigg(\int_{y_{j-\frac{1}{2}}}^{y_{j+\frac{1}{2}}} u(x^+_{i-\frac{1}{2}},y) - f(x^+_{i-\frac{1}{2}},y)dy\bigg)\\ =&\frac{6h_y}{h_x}\bar{q}^1_{ij} - \frac{2}{h_x}\bigg(\int_{y_{j-\frac{1}{2}}}^{y_{j+\frac{1}{2}}} q^1(x^-_{i+\frac{1}{2}},y)dy\bigg) - \frac{4}{h_x}\bigg(\int_{y_{j-\frac{1}{2}}}^{y_{j+\frac{1}{2}}} q^1(x^-_{i-\frac{1}{2}},y)dy\bigg) + q^2_{ij,3} - q^2_{i,j-1,3},\end{aligned}$$

and

(B.64)
$$\begin{aligned}&\frac{1}{\tau}\bigg(\int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} u(x,y^+_{j-\frac{1}{2}}) - f(x,y^+_{j-\frac{1}{2}})dx\bigg)\\ =&\frac{6h_x}{h_y}\bar{q}^2_{ij} - \frac{2}{h_y}\bigg(\int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} q^2(x,y^-_{j+\frac{1}{2}})dx\bigg) - \frac{4}{h_y}\bigg(\int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} q^2(x,y^-_{j-\frac{1}{2}})dx\bigg) + q^1_{ij,2} - q^1_{i-1,j,2},\end{aligned}$$

where $q_{ij,2}$ and $q_{ij,3}$ refer to the function values $q(x)$ at the bottom right and top left vertices of the $ij$-th rectangular cell. Then we need to get those point values. Specifically, when we consider the $i, j-1$-th rectangular cell, we will add a comma between i and j-1 to distinguish it from multiplication. Take $w = \delta_{1,i}(x)\delta_{-1,j}(y)$ and $\zeta = \delta_{-1,i}(x)\delta_{1,j}(y)$ in (2.39b) and (2.39c) individually, we have

$$q^1_{ij,2} = -\frac{6}{h_x^2}\int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} u(x,y^+_{j-\frac{1}{2}})dx + \frac{4}{h_x}u_{i+1,j,1} + \frac{2}{h_x}u_{ij,1}, \tag{B.65}$$

and

$$q^2_{ij,3} = -\frac{6}{(h_y)^2}\int_{y_{j-\frac{1}{2}}}^{y_{j+\frac{1}{2}}} u(x^+_{i-\frac{1}{2}},y)dy + \frac{4}{h_y}u_{i,j+1,1} + \frac{2}{h_y}u_{ij,1}, \tag{B.66}$$

where $u_{ij,1}$ means that the function value of $u(x)$ at the bottom left vertex of the $ij$-th rectangular cell. Then we need the representation of $u_{ij,1}$. Take $v = \delta_{-1,i}(x)\delta_{-1,j}(y)$ in (2.39a), then we have

$$\begin{aligned}\frac{1}{\tau}(u_{ij,1} - f_{ij,1}) =&\frac{6}{h^2}\int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} q^1(x,y^+_{j-\frac{1}{2}})dx + \frac{6}{(h_y)^2}\int_{y_{j-\frac{1}{2}}}^{y_{j+\frac{1}{2}}} q^2(x^+_{i-\frac{1}{2}},y)dy\\ &- \frac{2}{h_x}q^1_{ij,2} - \frac{4}{h_x}q^1_{i-1,j,2} - \frac{2}{h_y}q^2_{ij,3} - \frac{4}{h_y}q^2_{i,j-1,3}.\end{aligned} \tag{B.67}$$

To get $q^1_x$ and $q^2_y$, take $w = \delta_{-1,j}(y)$ and $\zeta = \delta_{-1,i}(x)$ separately, we have

$$\int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} q^1(x,y^+_{j-\frac{1}{2}})dx = u_{i+1,j,1} - u_{ij,1}, \tag{B.68}$$

and

$$\int_{y_{j-\frac{1}{2}}}^{y_{j+\frac{1}{2}}} q^2(x^+_{i-\frac{1}{2}},y)dy = u_{i,j+1,1} - u_{ij,1}. \tag{B.69}$$

We also need $\bar{q}^1$ and $\bar{q}^2$. Take $w = 1$ and $\zeta = 1$, we get

$$h_x h_y \bar{q}^1_{ij} = \int_{y_{j-\frac{1}{2}}}^{y_{j+\frac{1}{2}}} u(x^+_{i+\frac{1}{2}}, y)dy - \int_{y_{j-\frac{1}{2}}}^{y_{j+\frac{1}{2}}} u(x^+_{i-\frac{1}{2}}, y)dy, \tag{B.70}$$

and

$$h_x h_y \bar{q}^2_{ij} = \int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} u(x, y^+_{j+\frac{1}{2}})dy - \int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} u(x, y^+_{j-\frac{1}{2}})dy. \tag{B.71}$$

Now we have the same number of equations and variables and can get the formula of average values from the equations above from (B.60) to (B.71). First, we rewrite them into matrix form by tensor product.

(B.72a)
$$\frac{h_x h_y}{\tau}(\mathbf{I}_{N_y} \bigotimes \mathbf{I}_{N_x})(\bar{\mathbf{u}} - \bar{\mathbf{f}}) = (\mathbf{I}_{N_y} \bigotimes \mathbf{A}_{N_x})\mathbf{q}^1_{x^-} + (\mathbf{A}_{N_y} \bigotimes \mathbf{I}_{N_x})\mathbf{q}^2_{y^-},$$

(B.72b)
$$\mathbf{q}^1_{x^-} = -\frac{6h_y}{h_x}(\mathbf{I}_{N_y} \bigotimes \mathbf{I}_{N_x})\bar{\mathbf{u}} + \frac{1}{h_x}(\mathbf{I}_{N_y} \bigotimes \mathbf{B}_{N_x})\mathbf{u}_{x^+},$$

(B.72c)
$$\mathbf{q}^2_{y^-} = -\frac{6h_x}{h_y}(\mathbf{I}_{N_y} \bigotimes \mathbf{I}_{N_x})\bar{\mathbf{u}} + \frac{1}{h_y}(\mathbf{B}_{N_y} \bigotimes \mathbf{I}_{N_x})\mathbf{u}_{y^+},$$

(B.72d)
$$\frac{1}{\tau}\mathbf{u}_{x^+} = \frac{1}{\tau}\mathbf{f}_{x^+} + \frac{6h_y}{h_x}(\mathbf{I}_{N_y} \bigotimes \mathbf{I}_{N_x})\bar{\mathbf{q}}^1 - \frac{1}{h_x}(\mathbf{I}_{N_y} \bigotimes \mathbf{B}^\top_{N_x})\mathbf{q}^1_{x^-} + (\mathbf{A}_{N_y} \bigotimes \mathbf{I}_{N_x})\mathbf{q}^2_3,$$

(B.72e)
$$\frac{1}{\tau}\mathbf{u}_{y^+} = \frac{1}{\tau}\mathbf{f}_{y^+} + \frac{6h_x}{h_y}(\mathbf{I}_{N_y} \bigotimes \mathbf{I}_{N_x})\bar{\mathbf{q}}^2 - \frac{1}{h_y}(\mathbf{B}^\top_{N_y} \bigotimes \mathbf{I}_{N_x})\mathbf{q}^2_{y^-} + (\mathbf{I}_{N_y} \bigotimes \mathbf{A}_{N_x})\mathbf{q}^1_2,$$

(B.72f)
$$\mathbf{q}^1_2 = -\frac{6}{h_x^2}(\mathbf{I}_{N_y} \bigotimes \mathbf{I}_{N_x})\mathbf{u}_{y^+} + \frac{1}{h_x}(\mathbf{I}_{N_y} \bigotimes \mathbf{B}_{N_x})\mathbf{u}_1,$$

(B.72g)
$$\mathbf{q}^2_3 = -\frac{6}{h_y^2}(\mathbf{I}_{N_y} \bigotimes \mathbf{I}_{N_x})\mathbf{u}_{x^+} + \frac{1}{h_y}(\mathbf{B}_{N_y} \bigotimes \mathbf{I}_{N_x})\mathbf{u}_1,$$

(B.72h)
$$\frac{1}{\tau}\mathbf{u}_1 = \frac{1}{\tau}\mathbf{f}_1 + \frac{6}{h_x^2}(\mathbf{I}_{N_y} \bigotimes \mathbf{I}_{N_x})\mathbf{q}^1_{y^+} + \frac{6}{h_y^2}(\mathbf{I}_{N_y} \bigotimes \mathbf{I}_{N_x})\mathbf{q}^2_{x^+} - \frac{1}{h_x}(\mathbf{I}_{N_y} \bigotimes \mathbf{B}^\top_{N_x}) - \frac{1}{h_y}(\mathbf{B}^\top_{N_y} \bigotimes \mathbf{I}_{N_x})\mathbf{q}^2_3,$$

(B.72i)
$$\mathbf{q}^1_{y^+} = -(\mathbf{I}_{N_y} \bigotimes \mathbf{A}^\top_{N_x})\mathbf{u}_1,$$

(B.72j)
$$\mathbf{q}^2_{x^+} = -(\mathbf{A}^\top_{N_y} \bigotimes \mathbf{I}_{N_x})\mathbf{u}_1,$$

(B.72k)
$$h_x h_y \bar{\mathbf{q}}^1 = -(\mathbf{I}_{N_y} \bigotimes \mathbf{A}^\top_{N_x})\mathbf{u}_{x^+},$$

(B.72l)
$h_x h_y \bar{\mathbf{q}}^2 = -(\mathbf{A}_{N_y}^\top \bigotimes \mathbf{I}_{N_x})\mathbf{u}_{y^+}$.

Putting both (B.72b) and (B.72c) into (B.72a) and multiplying by $h_x h_y/\tau$, then we have

$$\mathbf{T}_3\bar{\mathbf{u}} = \mathbf{T}_4^1\frac{1}{h_x}\mathbf{u}_{x^+} + \mathbf{T}_4^2\frac{1}{h_y}\mathbf{u}_{y^+} + \bar{\mathbf{f}}. \tag{B.73}$$

Putting (B.72f), (B.72g), (B.72i) and (B.72j) into (B.72h) and multiplying by $\tau$, we have

$$\mathbf{T}_1\mathbf{u}_1 = \mathbf{f}_1 + \mathbf{T}_2^1\frac{1}{h_x}\mathbf{u}_{y^+} + \mathbf{T}_2^2\frac{1}{h_y}\mathbf{u}_{x^+}. \tag{B.74}$$

Putting (B.72b), (B.72g) and (B.72k) into (B.72d) and multiplying by $\tau/h_y$, we have

$$\mathbf{T}_5\frac{1}{h_y}\mathbf{u}_{x^+} = \frac{1}{h_y}\mathbf{f}_{x^+} + \mathbf{T}_2^1\bar{\mathbf{u}} + \mathbf{T}_4^2\mathbf{u}_1. \tag{B.75}$$

Putting (B.72c), (B.72f) and (B.72l) into (B.72e) and multiplying by $\tau/h$, we have

$$\mathbf{T}_6\frac{1}{h_x}\mathbf{u}_{y^+} = \frac{1}{h_x}\mathbf{f}_{y^+} + \mathbf{T}_2^2\bar{\mathbf{u}} + \mathbf{T}_4^1\mathbf{u}_1. \tag{B.76}$$

Using (B.74) to eliminate the $\mathbf{u}_1$ in (B.75) and (B.76), then we have

$$\mathbf{Q}\frac{1}{h_y}\mathbf{u}_{x^+} = \mathbf{P}_2\mathbf{T}_1\frac{1}{h_y}\mathbf{f}_{x^+} + \mathbf{T}_4^2\mathbf{T}_2^1\mathbf{T}_1\frac{1}{h_x}\mathbf{f}_{y^+} + \mathbf{P}_3\bar{\mathbf{u}} + \mathbf{P}_5\mathbf{f}_1, \tag{B.77}$$

and

$$\mathbf{Q}\frac{1}{h_x}\mathbf{u}_{y^+} = \mathbf{P}_1\mathbf{T}_1\frac{1}{h_x}\mathbf{f}_{y^+} + \mathbf{T}_4^1\mathbf{T}_2^2\mathbf{T}_1\frac{1}{h_y}\mathbf{f}_{x^+} + \mathbf{P}_4\bar{\mathbf{u}} + \mathbf{P}_6\mathbf{f}_1. \tag{B.78}$$

Collecting (B.77), (B.78) and (B.73) and multiplying it by $\mathbf{T}_1^{-1}$, we get our desired results directly.

B.7. **Proof of (i) in Theorem 2.9.**

**Proof:** Notice that

$$\mathbf{Q} = \mathbf{P}_1\mathbf{P}_2 - \mathbf{T}_4^1\mathbf{T}_4^2\mathbf{T}_2^1\mathbf{T}_2^2 = \mathbf{T}_1\Big(\mathbf{T}_1\mathbf{T}_5\mathbf{T}_6 - \mathbf{T}_2^2\mathbf{T}_4^2\mathbf{T}_6 - \mathbf{T}_2^1\mathbf{T}_4^1\mathbf{T}_5\Big),$$

$$\mathbf{P}_3 = \mathbf{T}_1\Big(\mathbf{P}_2\mathbf{T}_2^1 + \mathbf{T}_4^2\mathbf{T}_2^1\mathbf{T}_2^2\Big),\ \mathbf{P}_4 = \mathbf{T}_1\Big(\mathbf{P}_1\mathbf{T}_2^2 + \mathbf{T}_4^1\mathbf{T}_2^1\mathbf{T}_2^2\Big),$$

and we have

$$\mathbf{T}_1 = \mathbf{I}_{N_y}\bigotimes\mathbf{I}_{N_x} + \sigma_x\Big(\mathbf{I}_{N_y}\bigotimes\mathbf{T}_{1,x}\Big) + \sigma_y\Big(\mathbf{T}_{1,y}\bigotimes\mathbf{I}_{N_x}\Big)$$

$$= \mathbf{I}_{N_y}\bigotimes\mathbf{I}_{N_x} + \sigma_x\Big(\mathbf{I}_{N_y}\bigotimes\Big(26\mathbf{I}_{N_x} + 2\mathbf{G}_{N_x} + 8\mathbf{G}_{N_x}^{-1}\Big)\Big) + \sigma_y\Big(\Big(26\mathbf{I}_{N_y} + 2\mathbf{G}_{N_y} + 8\mathbf{G}_{N_y}^{-1}\Big)\bigotimes\mathbf{I}_{N_x}\Big),$$

therefore, $\mathbf{T}_1$ is strictly diagonally dominate and invertible. Then we have

(B.79)

$$\begin{aligned}\mathbf{F} =& \mathbf{Q}_4\mathbf{T}_3 - \mathbf{T}_1^{-1}\mathbf{T}_4\mathbf{P}_3 - \mathbf{T}_1^{-1}\mathbf{P}_4\mathbf{T}_4^2\\ =& \mathbf{T}_3\Big(\mathbf{T}_1\mathbf{T}_5\mathbf{T}_6 - \mathbf{T}_2^2\mathbf{T}_4^2\mathbf{T}_6 - \mathbf{T}_2^1\mathbf{T}_4^1\mathbf{T}_5\Big) - \mathbf{T}_4^1\Big(\mathbf{P}_2\mathbf{T}_2^1 + \mathbf{T}_4^2\mathbf{T}_2^1\mathbf{T}_2^2\Big) - \mathbf{T}_4^2\Big(\mathbf{P}_1\mathbf{T}_2^2 + \mathbf{T}_4^1\mathbf{T}_2^1\mathbf{T}_2^2\Big).\end{aligned}$$

Firstly, we calculate

$$\mathbf{Q}_4 = \mathbf{T}_1\mathbf{T}_5\mathbf{T}_6 - \mathbf{T}_2^2\mathbf{T}_4^2\mathbf{T}_6 - \mathbf{T}_2^1\mathbf{T}_4^1\mathbf{T}_5.$$

By definition we have

$$\begin{aligned}
\mathbf{T}_1\mathbf{T}_5\mathbf{T}_6 =& \left(\mathbf{I}_{N_y}\bigotimes\mathbf{I}_{N_x} + \sigma_x(\mathbf{I}_{N_y}\bigotimes\mathbf{T}_{1,x}) + \sigma_y(\mathbf{T}_{1,y}\bigotimes\mathbf{I}_{N_x})\right)\\
&\cdot\left(\mathbf{I}_{N_y}\bigotimes\mathbf{I}_{N_x} + \sigma_x\mathbf{I}_{N_y}\bigotimes\mathbf{T}_{1,x} + 6\sigma_y\mathbf{A}_{N_y}\bigotimes\mathbf{I}_{N_x}\right)\cdot\left(\mathbf{I}_{N_y}\bigotimes\mathbf{I}_{N_x} + \sigma_y\mathbf{T}_{1,y}\bigotimes\mathbf{I}_{N_x} + 6\sigma_x\mathbf{I}_{N_y}\bigotimes\mathbf{A}_{N}\right.\\
=&\mathbf{I}_{N_y}\bigotimes\mathbf{I}_{N_x} + \sigma_x\left(\mathbf{I}_{N_y}\bigotimes(6\mathbf{I}_{N_x} + \mathbf{T}_{1,x}) + \mathbf{I}_{N_y}\bigotimes\mathbf{T}_{1,x}\right) + \sigma_y\left((6\mathbf{A}_{N_y} + \mathbf{T}_{1,y})\bigotimes\mathbf{I}_{N_x} + \mathbf{T}_{1,y}\bigotimes\mathbf{I}_{N_x}\right)\\
&+ \sigma_x^2\left(\mathbf{I}_{N_y}\bigotimes 12\mathbf{T}_{1,x}\mathbf{A}_{N_x} + \mathbf{I}_{N_y}\bigotimes\mathbf{T}_{1,x}\mathbf{T}_{1,x}\right) + \sigma_y^2\left(12\mathbf{T}_{1,y}\mathbf{A}_{N_y}\bigotimes\mathbf{I}_{N_x} + \mathbf{T}_{1,y}\mathbf{T}_{1,y}\bigotimes\mathbf{I}_{N_x}\right)\\
&+ \sigma_x\sigma_y\left(3\mathbf{T}_{1,y}\bigotimes\mathbf{T}_{1,x} + 36\mathbf{A}_{N_y}\bigotimes\mathbf{A}_{N_x} + 6\mathbf{A}_{N_y}\bigotimes\mathbf{T}_{1,x} + 6\mathbf{T}_{1,y}\bigotimes\mathbf{A}_{N_x}\right)\\
&+ \sigma_x^2\sigma_y\left(\mathbf{T}_{1,y}\bigotimes\mathbf{T}_{1,x}\mathbf{T}_{1,x} + 36\mathbf{A}_{N_y}\bigotimes\mathbf{A}_{N_x}\mathbf{T}_{1,x} + 6\mathbf{T}_{1,y}\bigotimes\mathbf{T}_{1,x}\mathbf{A}_{N_x}\right)\\
&+ \sigma_x\sigma_y^2\left(\mathbf{T}_{1,y}\mathbf{T}_{1,y}\bigotimes\mathbf{T}_{1,x} + 36\mathbf{T}_{1,y}\mathbf{A}_{N_y}\bigotimes\mathbf{A}_{N_x} + 6\mathbf{T}_{1,y}\mathbf{A}_{N_y}\bigotimes\mathbf{T}_{1,x}\right)\\
&+ \sigma_x^3\left(6\mathbf{I}_{N_y}\bigotimes\mathbf{T}_{1,x}\mathbf{T}_{1,x}\mathbf{A}_{N_x}\right) + \sigma_y^3\left(6\mathbf{T}_{1,y}\mathbf{T}_{1,y}\mathbf{A}_{N_y}\bigotimes\mathbf{I}_{N_x}\right)
\end{aligned}$$

Similarly we have

$$\begin{aligned}
\mathbf{T}_2^2\mathbf{T}_4^2\mathbf{T}_6 &= 6\sigma_y^2\left(\mathbf{A}_{N_y}\mathbf{B}_{N_y}^\top\mathbf{B}_{N_y}\bigotimes(\mathbf{I}_{N_x} + 6\sigma_x\mathbf{A}_{N_x}) + \sigma_y\left(\mathbf{T}_{1,y}\mathbf{A}_{N_y}\mathbf{B}_{N_y}^\top\mathbf{B}_{N_y}\bigotimes\mathbf{I}_{N_x}\right)\right),\\
\mathbf{T}_2^1\mathbf{T}_4^1\mathbf{T}_5 &= 6\sigma_x^2\left(\ (\mathbf{I}_{N_y} + 6\sigma_y\mathbf{A}_{N_y})\bigotimes\mathbf{A}_{N_x}\mathbf{B}_{N_x}^\top\mathbf{B}_{N_x} + \sigma_x\left(\mathbf{I}_{N_y}\bigotimes\mathbf{T}_{1,x}\mathbf{A}_{N_x}\mathbf{B}_{N_x}^\top\mathbf{B}_{N_x}\right)\right).
\end{aligned}$$

Then we get

$$\begin{aligned}
\mathbf{Q}_4 =&\widetilde{\mathbf{Q}}_1 + \sigma_x\widetilde{\mathbf{Q}}_x + \sigma_y\widetilde{\mathbf{Q}}_y + \sigma_x^2\widetilde{\mathbf{Q}}_{xx} + \sigma_y^2\widetilde{\mathbf{Q}}_{yy} + \sigma_x\sigma_y\widetilde{\mathbf{Q}}_{xy}\\
&+ \sigma_x^2\sigma_y\widetilde{\mathbf{Q}}_{xxy} + \sigma_x\sigma_y^2\widetilde{\mathbf{Q}}_{xyy} + \sigma_x^3\widetilde{\mathbf{Q}}_{xxx} + \sigma_y^3\widetilde{\mathbf{Q}}_{yyy},
\end{aligned}$$

where

$$\begin{aligned}
\widetilde{\mathbf{Q}}_1 =&\mathbf{I}_{N_y}\bigotimes\mathbf{I}_{N_x},\\
\widetilde{\mathbf{Q}}_x =&\mathbf{I}_{N_y}\bigotimes\left(58\mathbf{I}_{N_x} + 10\mathbf{G}_{N_x}^{-1} + 4\mathbf{G}_{N_x}\right),\\
\widetilde{\mathbf{Q}}_y =&\left(58\mathbf{I}_{N_y} + 10\mathbf{G}_{N_y}^{-1} + 4\mathbf{G}_{N_y}\right)\bigotimes\mathbf{I}_{N_x},\\
\widetilde{\mathbf{Q}}_{xx} =&\mathbf{I}_{N_y}\bigotimes\left(924\mathbf{I}_{N_x} + 272\mathbf{G}_{N_x}^{-1} + 80\mathbf{G}_{N_x} + 4\mathbf{G}_{N_x}^2 + 16\mathbf{G}_{N_x}^{-2}\right),\\
\widetilde{\mathbf{Q}}_{yy} =&\left(924\mathbf{I}_{N_y} + 272\mathbf{G}_{N_y}^{-1} + 80\mathbf{G}_{N_y} + 4\mathbf{G}_{N_y}^2 + 16\mathbf{G}_{N_y}^{-2}\right)\bigotimes\mathbf{I}_{N_x},
\end{aligned}$$

$$
\begin{aligned}
\widetilde{\mathbf{Q}}_{xy} =&\mathbf{I}_{N_y}\bigotimes\Big(2376\mathbf{I}_{N_x}+480\mathbf{G}_{N_x}^{-1}+168\mathbf{G}_{N_x}\Big)+\mathbf{G}_{N_y}\bigotimes\Big(168\mathbf{I}_{N_x}+36\mathbf{G}_{N_x}^{-1}+12\mathbf{G}_{N_x}\Big)\\
&+\mathbf{G}_{N_y}^{-1}\bigotimes\Big(480\mathbf{I}_{N_x}+132\mathbf{G}_{N_x}^{-1}+36\mathbf{G}_{N_x}\Big),\\
\widetilde{\mathbf{Q}}_{xxy} =&\mathbf{I}_{N_y}\bigotimes\Big(22584\mathbf{I}_{N_x}+7792\mathbf{G}_{N_x}^{-1}+2800\mathbf{G}_{N_x}+104\mathbf{G}_{N_x}^{2}+416\mathbf{G}_{N_x}^{-2}\Big)\\
&+\mathbf{G}_{N_y}\bigotimes\Big(1704\mathbf{I}_{N_x}+616\mathbf{G}_{N_x}^{-1}+232\mathbf{G}_{N_x}+8\mathbf{G}_{N_x}^{2}+32\mathbf{G}_{N_x}^{-2}\Big)\\
&+\mathbf{G}_{N_y}^{-1}\bigotimes\Big(6384\mathbf{I}_{N_x}+2680\mathbf{G}_{N_x}^{-1}+1144\mathbf{G}_{N_x}+32\mathbf{G}_{N_x}^{2}+128\mathbf{G}_{N_x}^{-2}\Big),\\
\widetilde{\mathbf{Q}}_{xyy} =&\mathbf{I}_{N_y}\bigotimes\Big(22584\mathbf{I}_{N_x}+6384\mathbf{G}_{N_x}^{-1}+1704\mathbf{G}_{N_x}\Big)+\mathbf{G}_{N_y}\bigotimes\Big(2800\mathbf{I}_{N_x}+1144\mathbf{G}_{N_x}^{-1}+232\mathbf{G}_{N_x}\Big)\\
&+\mathbf{G}_{N_y}^{-1}\bigotimes\Big(7792\mathbf{I}_{N_x}+2680\mathbf{G}_{N_x}^{-1}+616\mathbf{G}_{N_x}\Big)+\mathbf{G}_{N_y}^{-2}\bigotimes\Big(416\mathbf{I}_{N_x}+128\mathbf{G}_{N_x}^{-1}+32\mathbf{G}_{N_x}\Big)\\
&+\mathbf{G}_{N_y}^{2}\bigotimes\Big(104\mathbf{I}_{N_x}+32\mathbf{G}_{N_x}^{-1}+8\mathbf{G}_{N_x}\Big),\\
\widetilde{\mathbf{Q}}_{xxx} =&\mathbf{I}_{N_y}\bigotimes\Big(1512\mathbf{I}_{N_x}-360\mathbf{G}_{N_x}^{-1}-792\mathbf{G}_{N_x}-72\mathbf{G}_{N_x}^{2}-288\mathbf{G}_{N_x}^{-2}\Big),\\
\widetilde{\mathbf{Q}}_{yyy} =&\Big(1512\mathbf{I}_{N_y}-360\mathbf{G}_{N_y}^{-1}-792\mathbf{G}_{N_y}-72\mathbf{G}_{N_y}^{2}-288\mathbf{G}_{N_y}^{-2}\Big)\bigotimes\mathbf{I}_{N_x}.
\end{aligned}
$$

Then we have

$$
\begin{aligned}
\mathbf{Q}_4\mathbf{T}_3 =&\widetilde{\mathbf{Q}}_1+\sigma_x\Big(\widetilde{\mathbf{Q}}_x+(\mathbf{I}_{N_y}\bigotimes 6\mathbf{A}_{N_x})\widetilde{\mathbf{Q}}_1\Big)+\sigma_y\Big(\widetilde{\mathbf{Q}}_y+(6\mathbf{A}_{N_y}\bigotimes\mathbf{I}_{N_x})\widetilde{\mathbf{Q}}_1\Big)\\
&+\sigma_x^2\Big(\widetilde{\mathbf{Q}}_{xx}+(\mathbf{I}_{N_y}\bigotimes 6\mathbf{A}_{N_x})\widetilde{\mathbf{Q}}_x\Big)+\sigma_y^2\Big(\widetilde{\mathbf{Q}}_{yy}+(6\mathbf{A}_{N_y}\bigotimes\mathbf{I}_{N_x})\widetilde{\mathbf{Q}}_y\Big)\\
&+\sigma_x\sigma_y\Big(\widetilde{\mathbf{Q}}_{xy}+\widetilde{\mathbf{Q}}_x(\mathbf{I}_{N_y}\bigotimes 6\mathbf{A}_{N_x})+\widetilde{\mathbf{Q}}_y(6\mathbf{A}_{N_y}\bigotimes\mathbf{I}_{N_x})\Big)\\
&+\sigma_x^2\sigma_y\Big(\widetilde{\mathbf{Q}}_{xxy}+(\mathbf{I}_{N_y}\bigotimes 6\mathbf{A}_{N_x})\widetilde{\mathbf{Q}}_{xy}+(6\mathbf{A}_{N_y}\bigotimes\mathbf{I}_{N_x})\widetilde{\mathbf{Q}}_{xx}\Big)\\
&+\sigma_x\sigma_y^2\Big(\widetilde{\mathbf{Q}}_{xyy}+(\mathbf{I}_{N_y}\bigotimes 6\mathbf{A}_{N_x})\widetilde{\mathbf{Q}}_{yy}+(6\mathbf{A}_{N_y}\bigotimes\mathbf{I}_{N_x})\widetilde{\mathbf{Q}}_{xy}\Big)\\
&+\sigma_x^3\Big(\widetilde{\mathbf{Q}}_{xxx}+\widetilde{\mathbf{Q}}_{xx}(\mathbf{I}_{N_y}\bigotimes 6\mathbf{A}_{N_x})\Big)+\sigma_y^3\Big(\widetilde{\mathbf{Q}}_{yyy}+\widetilde{\mathbf{Q}}_{yy}(6\mathbf{A}_{N_y}\bigotimes\mathbf{I}_{N_x})\Big)\\
&+\sigma_x^2\sigma_y\Big(\widetilde{\mathbf{Q}}_{xxy}+(\mathbf{I}_{N_y}\bigotimes 6\mathbf{A}_{N_x})\widetilde{\mathbf{Q}}_{xy}+(6\mathbf{A}_{N_y}\bigotimes\mathbf{I}_{N_x})\widetilde{\mathbf{Q}}_{xx}\Big)\\
&+\sigma_x\sigma_y^2\Big(\widetilde{\mathbf{Q}}_{xyy}+(\mathbf{I}_{N_y}\bigotimes 6\mathbf{A}_{N_x})\widetilde{\mathbf{Q}}_{yy}+(6\mathbf{A}_{N_y}\bigotimes\mathbf{I}_{N_x})\widetilde{\mathbf{Q}}_{xy}\Big)\\
&+\sigma_x^2\sigma_y^2\Big((\mathbf{I}_{N_y}\bigotimes 6\mathbf{A}_{N_x})\widetilde{\mathbf{Q}}_{xyy}+(6\mathbf{A}_{N_y}\bigotimes\mathbf{I}_{N_x})\widetilde{\mathbf{Q}}_{xxy}\Big)
\end{aligned}
$$

$$
\begin{aligned}
&+\sigma_x\sigma_y^3\bigg((\mathbf{I}_{N_y}\bigotimes 6\mathbf{A}_{N_x})\widetilde{\mathbf{Q}}_{yyy}+(6\mathbf{A}_{N_y}\bigotimes\mathbf{I}_{N_x})\widetilde{\mathbf{Q}}_{xyy}\bigg)\\
&+\sigma_x^3\sigma_y\bigg((\mathbf{I}_{N_y}\bigotimes 6\mathbf{A}_{N_x})\widetilde{\mathbf{Q}}_{xxy}+(6\mathbf{A}_{N_y}\bigotimes\mathbf{I}_{N_x})\widetilde{\mathbf{Q}}_{xxx}\bigg)\\
&+\sigma_x^4\widetilde{\mathbf{Q}}_{xxx}(\mathbf{I}_{N_y}\bigotimes 6\mathbf{A}_{N_x})+\sigma_y^4\widetilde{\mathbf{Q}}_{yyy}(6\mathbf{A}_{N_y}\bigotimes\mathbf{I}_{N_x}).
\end{aligned}
$$

Similarly by definition we have

$$
\begin{aligned}
\mathbf{T}_1^{-1}\mathbf{T}_4^1\mathbf{P}_3=&6\sigma_x^2\bigg(\mathbf{I}_{N_y}\bigotimes\mathbf{A}_{N_x}\mathbf{B}_{N_x}^\top\mathbf{B}_{N_x}\bigg)+6\sigma_x^3\bigg(\mathbf{I}_{N_y}\bigotimes\mathbf{A}_{N_x}\mathbf{B}_{N_x}^\top\mathbf{B}_{N_x}(\mathbf{T}_{1,x}+6\mathbf{A}_{N_x})\bigg)\\
&+12\sigma_x^2\sigma_y\bigg(\mathbf{T}_{1,y}\bigotimes\mathbf{A}_{N_x}\mathbf{B}_{N_x}^\top\mathbf{B}_{N_x}\bigg)+6\sigma_x^3\sigma_y\bigg(\mathbf{T}_{1,y}\bigotimes\mathbf{A}_{N_x}\mathbf{B}_{N_x}^\top\mathbf{B}_{N_x}(\mathbf{T}_{1,x}+6\mathbf{A}_{N_x})\bigg)\\
&+6\sigma_x^4\bigg(\mathbf{I}_{N_y}\bigotimes\mathbf{A}_{N_x}\mathbf{B}_{N_x}^\top\mathbf{B}_{N_x}(6\mathbf{A}_{N_x}\mathbf{T}_{1,x}-6\mathbf{A}_{N_x}\mathbf{B}_{N_x}^\top\mathbf{B}_{N_x})\bigg)\\
&+6\sigma_x^2\sigma_y^2\bigg((\mathbf{T}_{1,y}\mathbf{T}_{1,y}+6\mathbf{A}_{N_y}\mathbf{B}_{N_y}^\top\mathbf{B}_{N_y})\bigotimes\mathbf{A}_{N_x}\mathbf{B}_{N_x}^\top\mathbf{B}_{N_x}\bigg),
\end{aligned}
$$

and

$$
\begin{aligned}
\mathbf{T}_1^{-1}\mathbf{T}_4^2\mathbf{P}_4=&6\sigma_y^2\bigg(\mathbf{A}_{N_y}\mathbf{B}_{N_y}^\top\mathbf{B}_{N_y}\bigotimes\mathbf{I}_{N_x}\bigg)+12\sigma_x\sigma_y^2\bigg(\mathbf{A}_{N_y}\mathbf{B}_{N_y}^\top\mathbf{B}_{N_y}\bigotimes\mathbf{T}_{1,x}\bigg)\\
&+6\sigma_y^3\bigg(\mathbf{A}_{N_y}\mathbf{B}_{N_y}^\top\mathbf{B}_{N_y}(\mathbf{T}_{1,y}+6\mathbf{A}_{N_y})\bigotimes\mathbf{I}_{N_x}\bigg)\\
&+6\sigma_x\sigma_y^3\bigg(\mathbf{A}_{N_y}\mathbf{B}_{N_y}^\top\mathbf{B}_{N_y}(\mathbf{T}_{1,y}+6\mathbf{A}_{N_y})\bigotimes\mathbf{T}_{1,x}\bigg)\\
&+6\sigma_x^2\sigma_y^2\bigg(\mathbf{A}_{N_y}\mathbf{B}_{N_y}^\top\mathbf{B}_{N_y}\bigotimes(\mathbf{T}_{1,x}\mathbf{T}_{1,x}+6\mathbf{A}_{N_x}\mathbf{B}_{N_x}^\top\mathbf{B}_{N_x})\bigg)\\
&+36\sigma_y^4\bigg(\mathbf{A}_{N_y}\mathbf{B}_{N_y}^\top\mathbf{B}_{N_y}(\mathbf{T}_{1,y}\mathbf{A}_{N_y}-\mathbf{A}_{N_y}\mathbf{B}_{N_y}^\top\mathbf{B}_{N_y})\bigotimes\mathbf{I}_{N_x}\bigg).
\end{aligned}
$$

Then we have

$$
\begin{aligned}
\mathbf{F}=&\widetilde{\mathbf{Q}}_1+\sigma_x\widetilde{\mathbf{F}}_x+\sigma_y\widetilde{\mathbf{F}}_y+\sigma_x^2\widetilde{\mathbf{F}}_{xx}+\sigma_y^2\widetilde{\mathbf{F}}_{yy}+\sigma_x\sigma_y\widetilde{\mathbf{F}}_{xy}+\sigma_x^3\widetilde{\mathbf{F}}_{xxx}+\sigma_y^3\widetilde{\mathbf{F}}_{yyy}\\
&+\sigma_x^2\sigma_y\widetilde{\mathbf{F}}_{xxy}+\sigma_x\sigma_y^2\widetilde{\mathbf{F}}_{xyy}+\sigma_x^2\sigma_y^2\widetilde{\mathbf{F}}_{xxyy}\\
&+\sigma_x\sigma_y^3\widetilde{\mathbf{F}}_{xyyy}+\sigma_x^3\sigma_y\widetilde{\mathbf{F}}_{xxxy}+\sigma_x^4\widetilde{\mathbf{F}}_{xxxx}+\sigma_y^4\widetilde{\mathbf{F}}_{yyyy},
\end{aligned}
\tag{B.80}
$$

where

$$
\begin{aligned}
\widetilde{\mathbf{F}}_x=&\widetilde{\mathbf{Q}}_x+(\mathbf{I}_{N_y}\bigotimes 6\mathbf{A}_{N_x})\widetilde{\mathbf{Q}}_1=\mathbf{I}_{N_y}\bigotimes\bigg(64\mathbf{I}_{N_x}+4\mathbf{G}_{N_x}+4\mathbf{G}_{N_x}^{-1}\bigg),\\
\widetilde{\mathbf{F}}_y=&\widetilde{\mathbf{Q}}_y+(6\mathbf{A}_{N_y}\bigotimes\mathbf{I}_{N_x})\widetilde{\mathbf{Q}}_1=\bigg(64\mathbf{I}_{N_y}+4\mathbf{G}_{N_y}+4\mathbf{G}_{N_y}^{-1}\bigg)\bigotimes\mathbf{I}_{N_x},\\
\widetilde{\mathbf{F}}_{xx}=&\widetilde{\mathbf{Q}}_{xx}+(\mathbf{I}_{N_y}\bigotimes 6\mathbf{A}_{N_x})\widetilde{\mathbf{Q}}_x-6\mathbf{I}_{N_y}\bigotimes\mathbf{A}_{N_x}\mathbf{B}_{N_x}^\top\mathbf{B}_{N_x}\\
=&\mathbf{I}_{N_y}\bigotimes\bigg(1176\mathbf{I}_{N_x}+56\mathbf{G}_{N_x}+56\mathbf{G}_{N_x}^{-1}+4\mathbf{G}_{N_x}^2+4\mathbf{G}_{N_x}^{-2}\bigg),
\end{aligned}
$$

$$
\begin{aligned}
\widetilde{\mathbf{F}}_{yy} =&\widetilde{\mathbf{Q}}_{yy} + (6\mathbf{A}_{N_y} \bigotimes \mathbf{I}_{N_x})\widetilde{\mathbf{Q}}_{y} - 6\mathbf{A}_{N_y}\mathbf{B}_{N_y}^{\top}\mathbf{B}_{N_y} \bigotimes \mathbf{I}_{N_x} \\
=&\left(1176\mathbf{I}_{N_y} + 56\mathbf{G}_{N_y} + 56\mathbf{G}_{N_y}^{-1} + 4\mathbf{G}_{N_y}^{2} + 4\mathbf{G}_{N_y}^{-2}\right) \bigotimes \mathbf{I}_{N_x}, \\
\widetilde{\mathbf{F}}_{xy} =&\widetilde{\mathbf{Q}}_{xy} + \widetilde{\mathbf{Q}}_{y}(\mathbf{I}_{N_y} \bigotimes 6\mathbf{A}_{N_x}) + \widetilde{\mathbf{Q}}_{x}(6\mathbf{A}_{N_y} \bigotimes \mathbf{I}_{N_x}) \\
=&\mathbf{I}_{N_y} \bigotimes \left(3072\mathbf{I}_{N_x} + 192\mathbf{G}_{N_x} + 192\mathbf{G}_{N_x}^{-1}\right) + \mathbf{G}_{N_y} \bigotimes \left(192\mathbf{I}_{N_x} + 12\mathbf{G}_{N_x} + 12\mathbf{G}_{N_x}^{-1}\right) \\
&+ \mathbf{G}_{N_y}^{-1} \bigotimes \left(192\mathbf{I}_{N_x} + 12\mathbf{G}_{N_x} + 12\mathbf{G}_{N_x}^{-1}\right), \\
\widetilde{\mathbf{F}}_{xxx} =&\widetilde{\mathbf{Q}}_{xxx} + \widetilde{\mathbf{Q}}_{xx}(\mathbf{I}_{N_y} \bigotimes 6\mathbf{A}_{N_x}) - \mathbf{I}_{N_y} \bigotimes 6\mathbf{A}_{N_x}\mathbf{B}_{N_x}^{\top}\mathbf{B}_{N_x}(\mathbf{T}_{1,x} + 6\mathbf{A}_{N_x}), \\
=&\mathbf{I}_{N_y} \bigotimes \left(4320\mathbf{I}_{N_x} + -2016\mathbf{G}_{N_x} - 2016\mathbf{G}_{N_x}^{-1} - 144\mathbf{G}_{N_x}^{2} - 144\mathbf{G}_{N_x}^{-2}\right), \\
\widetilde{\mathbf{F}}_{yyy} =&\left(4320\mathbf{I}_{N_y} - 2016\mathbf{G}_{N_y} - 2016\mathbf{G}_{N_y}^{-1} - 144\mathbf{G}_{N_y}^{2} - 144\mathbf{G}_{N_y}^{-2}\right) \bigotimes \mathbf{I}_{N_x}, \\
\widetilde{\mathbf{F}}_{xxy} =&\widetilde{\mathbf{Q}}_{xxy} + \widetilde{\mathbf{Q}}_{xy}(\mathbf{I}_{N_y} \bigotimes 6\mathbf{A}_{N_x}) + \widetilde{\mathbf{Q}}_{xx}(6\mathbf{A}_{N_y} \bigotimes \mathbf{I}_{N_x}) - 12\mathbf{T}_{1,y} \bigotimes \mathbf{A}_{N_x}\mathbf{B}_{N_x}^{\top}\mathbf{B}_{N_x} \\
=&\mathbf{I}_{N_y} \bigotimes \left(37632\mathbf{I}_{N_x} + 1792\mathbf{G}_{N_x} + 1792\mathbf{G}_{N_x}^{-1} + 128\mathbf{G}_{N_x}^{2} + 128\mathbf{G}_{N_x}^{-2}\right) \\
&+ \mathbf{G}_{N_y} \bigotimes \left(2352\mathbf{I}_{N_x} + 112\mathbf{G}_{N_x} + 112\mathbf{G}_{N_x}^{-1} + 8\mathbf{G}_{N_x}^{2} + 8\mathbf{G}_{N_x}^{-2}\right) \\
&+ \mathbf{G}_{N_y}^{-1} \bigotimes \left(2352\mathbf{I}_{N_x} + 112\mathbf{G}_{N_x} + 112\mathbf{G}_{N_x}^{-1} + 8\mathbf{G}_{N_x}^{2} + 8\mathbf{G}_{N_x}^{-2}\right), \\
\widetilde{\mathbf{F}}_{xyy} =&\widetilde{\mathbf{Q}}_{xyy} + \widetilde{\mathbf{Q}}_{yy}(\mathbf{I}_{N_y} \bigotimes 6\mathbf{A}_{N_x}) + \widetilde{\mathbf{Q}}_{xy}(6\mathbf{A}_{N_y} \bigotimes \mathbf{I}_{N_x}) - 12\mathbf{A}_{N_y}\mathbf{B}_{N_y}^{\top}\mathbf{B}_{N_y} \bigotimes \mathbf{T}_{1,x} \\
=&\mathbf{I}_{N_y} \bigotimes \left(37632\mathbf{I}_{N_x} + 2352\mathbf{G}_{N_x} + 2352\mathbf{G}_{N_x}^{-1}\right) + \mathbf{G}_{N_y} \bigotimes \left(1792\mathbf{I}_{N_x} + 112\mathbf{G}_{N_x} + 112\mathbf{G}_{N_x}^{-1}\right) \\
&+ \mathbf{G}_{N_y}^{2} \bigotimes \left(128\mathbf{I}_{N_x} + 8\mathbf{G}_{N_x} + 8\mathbf{G}_{N_x}^{-1}\right) + \mathbf{G}_{N_y}^{-2} \bigotimes \left(128\mathbf{I}_{N_x} + 8\mathbf{G}_{N_x} + 8\mathbf{G}_{N_x}^{-1}\right) \\
&+ \mathbf{G}_{N_y}^{-1} \bigotimes \left(1792\mathbf{I}_{N_x} + 112\mathbf{G}_{N_x} + 112\mathbf{G}_{N_x}^{-1}\right), \\
\widetilde{\mathbf{F}}_{xxyy} =&\widetilde{\mathbf{Q}}_{xyy}(\mathbf{I}_{N_y} \bigotimes 6\mathbf{A}_{N_x}) - (\mathbf{T}_{1,y}\mathbf{T}_{1,y} + 6\mathbf{A}_{N_y}\mathbf{B}_{N_y}^{\top}\mathbf{B}_{N_y}) \bigotimes 6\mathbf{A}_{N_x}\mathbf{B}_{N_x}^{\top}\mathbf{B}_{N_x} \\
&+ \widetilde{\mathbf{Q}}_{xxy}(6\mathbf{A}_{N_y} \bigotimes \mathbf{I}_{N_x}) - 6\mathbf{A}_{N_y}\mathbf{B}_{N_y}^{\top}\mathbf{B}_{N_y} \bigotimes (\mathbf{T}_{1,x}\mathbf{T}_{1,x} + 6\mathbf{A}_{N_x}\mathbf{B}_{N_x}^{\top}\mathbf{B}_{N_x}) \\
=&\mathbf{I}_{N_y} \bigotimes \left(138240\mathbf{I}_{N_x} - 22752\mathbf{G}_{N_x} - 22752\mathbf{G}_{N_x}^{-1} + 288\mathbf{G}_{N_x}^{2} + 288\mathbf{G}_{N_x}^{-2}\right) \\
&+ \mathbf{G}_{N_y} \bigotimes \left( - 22752\mathbf{I}_{N_x} - 11808\mathbf{G}_{N_x} - 11808\mathbf{G}_{N_x}^{-1} - 144\mathbf{G}_{N_x}^{2} - 144\mathbf{G}_{N_x}^{-2}\right) \\
&+ \mathbf{G}_{N_y}^{-1} \bigotimes \left( - 22752\mathbf{I}_{N_x} - 11808\mathbf{G}_{N_x} - 11808\mathbf{G}_{N_x}^{-1} - 144\mathbf{G}_{N_x}^{2} - 144\mathbf{G}_{N_x}^{-2}\right)
\end{aligned}
$$

$$+\mathbf{G}_{N_y}^2\bigotimes\left(288\mathbf{I}_{N_x}-144\mathbf{G}_{N_x}-144\mathbf{G}_{N_x}^{-1}\right)+\mathbf{G}_{N_y}^{-2}\bigotimes\left(288\mathbf{I}_{N_x}-144\mathbf{G}_{N_x}-144\mathbf{G}_{N_x}^{-1}\right),$$

$$\widetilde{\mathbf{F}}_{xxxx}=\mathbf{I}_{N_y}\bigotimes\left(7776\mathbf{I}_{N_x}-5184\mathbf{G}_{N_x}-5184\mathbf{G}_{N_x}^{-1}+1296\mathbf{G}_{N_x}^2+1296\mathbf{G}_{N_x}^{-2}\right)$$

$$\widetilde{\mathbf{F}}_{yyyy}=\left(7776\mathbf{I}_{N_y}-5184\mathbf{G}_{N_y}-5184\mathbf{G}_{N_y}^{-1}+1296\mathbf{G}_{N_y}^2+1296\mathbf{G}_{N_y}^{-2}\right)\bigotimes\mathbf{I}_{N_x}.$$

By collecting the same terms, we can get the exact formulas of each elements of $\mathbf{F}$ as follows. We will show the elements' formulas behind their location in the matrix.

$$\begin{aligned}
\mathbf{I}_{N_y}\bigotimes\mathbf{I}_{N_x}:&1+64\sigma_x+64\sigma_y+1176\sigma_x^2+3072\sigma_x\sigma_y+4320\sigma_x^3+37632\sigma_x^2\sigma_y+37632\sigma_x\sigma_y^2\\
&+138240\sigma_x^2\sigma_y^2+69120\sigma_x\sigma_y^3+69120\sigma_x^3\sigma_y+7776\sigma_x^4+7776\sigma_y^4,\\
\mathbf{I}_{N_y}\bigotimes\mathbf{G}_{N_x}:&4\sigma_x+56\sigma_x^2+192\sigma_x\sigma_y-2016\sigma_x^3+1792\sigma_x^2\sigma_y+2352\sigma_x\sigma_y^2-22752\sigma_x^2\sigma_y^2\\
&+4320\sigma_x\sigma_y^3-32256\sigma_x^3\sigma_y-5184\sigma_x^4,\\
\mathbf{I}_{N_y}\bigotimes\mathbf{G}_{N_x}^{-1}:&4\sigma_x+56\sigma_x^2+192\sigma_x\sigma_y-2016\sigma_x^3+1792\sigma_x^2\sigma_y+2352\sigma_x\sigma_y^2-22752\sigma_x^2\sigma_y^2\\
&+4320\sigma_x\sigma_y^3-32256\sigma_x^3\sigma_y-5184\sigma_x^4,\\
\mathbf{I}_{N_y}\bigotimes\mathbf{G}_{N_x}^2:&4\sigma_x^2-144\sigma_x^3-2304\sigma_x^3\sigma_y+128\sigma_x^2\sigma_y+288\sigma_x^2\sigma_y^2+1296\sigma_x^4,\\
\mathbf{I}_{N_y}\bigotimes\mathbf{G}_{N_x}^{-2}:&4\sigma_x^2-144\sigma_x^3-2304\sigma_x^3\sigma_y+128\sigma_x^2\sigma_y+288\sigma_x^2\sigma_y^2+1296\sigma_x^4,\\
\mathbf{G}_{N_y}\bigotimes\mathbf{I}_{N_x}:&4\sigma_y+56\sigma_y^2+192\sigma_x\sigma_y-2016\sigma_y^3+2352\sigma_x^2\sigma_y+1792\sigma_x\sigma_y^2-22752\sigma_x^2\sigma_y^2\\
&+4320\sigma_x^3\sigma_y-32256\sigma_x\sigma_y^3-5184\sigma_y^4,\\
\mathbf{G}_{N_y}\bigotimes\mathbf{G}_{N_x}:&12\sigma_x\sigma_y+112\sigma_x^2\sigma_y+112\sigma_x\sigma_y^2-11808\sigma_x^2\sigma_y^2-2016\sigma_x\sigma_y^3-2016\sigma_x^3\sigma_y,\\
\mathbf{G}_{N_y}\bigotimes\mathbf{G}_{N_x}^{-1}:&12\sigma_x\sigma_y+112\sigma_x^2\sigma_y+112\sigma_x\sigma_y^2-11808\sigma_x^2\sigma_y^2-2016\sigma_x\sigma_y^3-2016\sigma_x^3\sigma_y,\\
\mathbf{G}_{N_y}\bigotimes\mathbf{G}_{N_x}^2:&8\sigma_x^2\sigma_y-144\sigma_x^2\sigma_y^2-144\sigma_x^3\sigma_y,\\
\mathbf{G}_{N_y}\bigotimes\mathbf{G}_{N_x}^{-2}:&8\sigma_x^2\sigma_y-144\sigma_x^2\sigma_y^2-144\sigma_x^3\sigma_y,\\
\mathbf{G}_{N_y}^{-1}\bigotimes\mathbf{I}_{N_x}:&4\sigma_y+56\sigma_y^2+192\sigma_x\sigma_y-2016\sigma_y^3+2352\sigma_x^2\sigma_y+1792\sigma_x\sigma_y^2-22752\sigma_x^2\sigma_y^2\\
&+4320\sigma_x^3\sigma_y-32256\sigma_x\sigma_y^3-5184\sigma_y^4,\\
\mathbf{G}_{N_y}^{-1}\bigotimes\mathbf{G}_{N_x}:&12\sigma_x\sigma_y+112\sigma_x^2\sigma_y+112\sigma_x\sigma_y^2-11808\sigma_x^2\sigma_y^2-2016\sigma_x\sigma_y^3-2016\sigma_x^3\sigma_y,\\
\mathbf{G}_{N_y}^{-1}\bigotimes\mathbf{G}_{N_x}^{-1}:&12\sigma_x\sigma_y+112\sigma_x^2\sigma_y+112\sigma_x\sigma_y^2-11808\sigma_x^2\sigma_y^2-2016\sigma_x\sigma_y^3-2016\sigma_x^3\sigma_y,\\
\mathbf{G}_{N_y}^{-1}\bigotimes\mathbf{G}_{N_x}^2:&8\sigma_x^2\sigma_y-144\sigma_x^2\sigma_y^2-144\sigma_x^3\sigma_y,\\
\mathbf{G}_{N_y}^{-1}\bigotimes\mathbf{G}_{N_x}^{-2}:&8\sigma_x^2\sigma_y-144\sigma_x^2\sigma_y^2-144\sigma_x^3\sigma_y,\\
\mathbf{G}_{N_y}^2\bigotimes\mathbf{I}_{N_x}:&4\sigma_y^2-144\sigma_y^3+128\sigma_x\sigma_y^2+288\sigma_x^2\sigma_y^2-2304\sigma_x\sigma_y^3+1296\sigma_y^4,\\
\mathbf{G}_{N_y}^2\bigotimes\mathbf{G}_{N_x}:&8\sigma_x\sigma_y^2-144\sigma_x^2\sigma_y^2-144\sigma_x\sigma_y^3,
\end{aligned}$$

$$\mathbf{G}_{N_y}^{2}\bigotimes\mathbf{G}_{N_x}^{-1}:8\sigma_x\sigma_y^2-144\sigma_x^2\sigma_y^2-144\sigma_x\sigma_y^3,$$
$$\mathbf{G}_{N_y}^{-2}\bigotimes\mathbf{I}_{N_x}:4\sigma_y^2-144\sigma_y^3+128\sigma_x\sigma_y^2+288\sigma_x^2\sigma_y^2-2304\sigma_x\sigma_y^3+1296\sigma_y^4,$$
$$\mathbf{G}_{N_y}^{-2}\bigotimes\mathbf{G}_{N_x}:8\sigma_x\sigma_y^2-144\sigma_x^2\sigma_y^2-144\sigma_x\sigma_y^3,$$
$$\mathbf{G}_{N_y}^{-2}\bigotimes\mathbf{G}_{N_x}^{-1}:8\sigma_x\sigma_y^2-144\sigma_x^2\sigma_y^2-144\sigma_x\sigma_y^3.$$

The row sums of $\mathbf{F}$ are the same and are equal to

$$46656(\sigma_x\sigma_y^2+\sigma_x^2\sigma_y)+3888\sigma_x\sigma_y+72\sigma_x+72\sigma_y+1296\sigma_x^2+1296\sigma_y^2+1.$$

If we take $\sigma_x=\sigma_y\geq 1$ which is far away sharp, we can check easily that all of the off-diagonal elements of $\mathbf{F}$ are negative and the diagonal elements are positive, and $\mathbf{F}$ is strictly diagonally dominant, then we conclude that $\mathbf{F}$ is an $M$-matrix. □

B.8. **Proof of (ii) and (iii) in Theorem 2.9.** We will calculate the matrices $\mathbf{W}_\varrho$ ($\varrho=1,2,3,4$) directly by definition and prove that their elements are all non-negative and at least one positive. Notice that $\left\{\mathbf{G}_{N_x}^{i_x}\bigotimes\mathbf{G}_{N_y}^{i_y}\right\}\Big|_{i_x,i_y=1}^{Nx,Ny}$ can be viewed as a set of basis of $\mathcal{R}^{N_x^2\times N_y^2}$. Therefore, we only show that the non-zero elements of the defined matrix under the representation of this basis here. Firstly we calculate $\mathbf{Q}_\varrho$ ($\varrho=1,2,3,4$) and show their representation in the form (basis: coefficients) as below.
$\mathbf{Q}_1$:

$$\mathbf{I}_{N_y}\bigotimes\mathbf{I}_{N_x}:-2\sigma_x-60\sigma_x^2-104\sigma_x\sigma_y-1560\sigma_x\sigma_y^2-1560\sigma_x^2\sigma_y-216\sigma_x^3,$$
$$\mathbf{I}_{N_y}\bigotimes\mathbf{G}_{N_x}:4\sigma_x+124\sigma_x^2+208\sigma_x\sigma_y+3120\sigma_x\sigma_y^2+3224\sigma_x^2\sigma_y+360\sigma_x^3,$$
$$\mathbf{I}_{N_y}\bigotimes\mathbf{G}_{N_x}^{-1}:-2\sigma_x-68\sigma_x^2-104\sigma_x\sigma_y-1560\sigma_x\sigma_y^2-1768\sigma_x^2\sigma_y-72\sigma_x^3$$
$$\mathbf{I}_{N_y}\bigotimes\mathbf{G}_{N_x}^{2}:8\sigma_x^2+208\sigma_x^2\sigma_y-144\sigma_x^3,$$
$$\mathbf{I}_{N_y}\bigotimes\mathbf{G}_{N_x}^{-2}:-4\sigma_x^2-104\sigma_x^2\sigma_y+72\sigma_x^3,$$
$$\mathbf{G}_{N_y}\bigotimes\mathbf{I}_{N_x}:-8\sigma_x\sigma_y-304\sigma_x\sigma_y^2-120\sigma_x^2\sigma_y,$$
$$\mathbf{G}_{N_y}\bigotimes\mathbf{G}_{N_x}:608\sigma_x\sigma_y^2+248\sigma_x^2\sigma_y+16\sigma_x\sigma_y,$$
$$\mathbf{G}_{N_y}\bigotimes\mathbf{G}_{N_x}^{-1}:-8\sigma_x\sigma_y-304\sigma_x\sigma_y^2-136\sigma_x^2\sigma_y,$$
$$\mathbf{G}_{N_y}\bigotimes\mathbf{G}_{N_x}^{2}:16\sigma_x^2\sigma_y,$$
$$\mathbf{G}_{N_y}\bigotimes\mathbf{G}_{N_x}^{-2}:-8\sigma_x^2\sigma_y,$$
$$\mathbf{G}_{N_y}^{-1}\bigotimes\mathbf{I}_{N_x}:-32\sigma_x\sigma_y-688\sigma_x\sigma_y^2-480\sigma_x^2\sigma_y,$$
$$\mathbf{G}_{N_y}^{-1}\bigotimes\mathbf{G}_{N_x}:64\sigma_x\sigma_y+1376\sigma_x\sigma_y^2+992\sigma_x^2\sigma_y,$$
$$\mathbf{G}_{N_y}^{-1}\bigotimes\mathbf{G}_{N_x}^{-1}:-32\sigma_x\sigma_y-688\sigma_x\sigma_y^2-544\sigma_x^2\sigma_y,$$
$$\mathbf{G}_{N_y}^{-1}\bigotimes\mathbf{G}_{N_x}^{2}:64\sigma_x^2\sigma_y,$$

$$\mathbf{G}_{N_y}^{-1}\bigotimes\mathbf{G}_{N_x}^{-2}:-32\sigma_x^2\sigma_y,$$
$$\mathbf{G}_{N_y}^{2}\bigotimes\mathbf{I}_{N_x}:-8\sigma_x\sigma_y^2,$$
$$\mathbf{G}_{N_y}^{2}\bigotimes\mathbf{G}_{N_x}:16\sigma_x\sigma_y^2,$$
$$\mathbf{G}_{N_y}^{2}\bigotimes\mathbf{G}_{N_x}^{-1}:-8\sigma_x\sigma_y^2,$$
$$\mathbf{G}_{N_y}^{-2}\bigotimes\mathbf{I}_{N_x}:-32\sigma_x\sigma_y^2,$$
$$\mathbf{G}_{N_y}^{-2}\bigotimes\mathbf{G}_{N_x}:64\sigma_x\sigma_y^2,$$
$$\mathbf{G}_{N_y}^{-2}\bigotimes\mathbf{G}_{N_x}^{-1}:-32\sigma_x\sigma_y^2.$$

$\mathbf{Q}_2$:

$$\mathbf{I}_{N_y}\bigotimes\mathbf{I}_{N_x}:-2\sigma_y-60\sigma_y^2-104\sigma_x\sigma_y-1560\sigma_x\sigma_y^2-1560\sigma_x^2\sigma_y-216\sigma_y^3,$$
$$\mathbf{I}_{N_y}\bigotimes\mathbf{G}_{N_x}:-8\sigma_x\sigma_y-120\sigma_x\sigma_y^2-304\sigma_x^2\sigma_y,$$
$$\mathbf{I}_{N_y}\bigotimes\mathbf{G}_{N_x}^{-1}:-32\sigma_x\sigma_y-480\sigma_x\sigma_y^2-688\sigma_x^2\sigma_y,$$
$$\mathbf{I}_{N_y}\bigotimes\mathbf{G}_{N_x}^{2}:-8\sigma_x^2\sigma_y,$$
$$\mathbf{I}_{N_y}\bigotimes\mathbf{G}_{N_x}^{-2}:-32\sigma_x^2\sigma_y$$
$$\mathbf{G}_{N_y}\bigotimes\mathbf{I}_{N_x}:4\sigma_y+124\sigma_y^2+208\sigma_x\sigma_y+3224\sigma_x\sigma_y^2+3120\sigma_x^2\sigma_y+360\sigma_y^3,$$
$$\mathbf{G}_{N_y}\bigotimes\mathbf{G}_{N_x}:16\sigma_x\sigma_y+248\sigma_x\sigma_y^2+608\sigma_x^2\sigma_y,$$
$$\mathbf{G}_{N_y}\bigotimes\mathbf{G}_{N_x}^{-1}:1376\sigma_x^2\sigma_y+64\sigma_x\sigma_y+992\sigma_x\sigma_y^2,$$
$$\mathbf{G}_{N_y}\bigotimes\mathbf{G}_{N_x}^{2}:16\sigma_x^2\sigma_y,$$
$$\mathbf{G}_{N_y}\bigotimes\mathbf{G}_{N_x}^{-2}:64\sigma_x^2\sigma_y,$$
$$\mathbf{G}_{N_y}^{-1}\bigotimes\mathbf{I}_{N_x}:-2\sigma_y-104\sigma_x\sigma_y-1768\sigma_x\sigma_y^2-1560\sigma_x^2\sigma_y-68\sigma_y^2-72\sigma_y^3,$$
$$\mathbf{G}_{N_y}^{-1}\bigotimes\mathbf{G}_{N_x}:-8\sigma_x\sigma_y-136\sigma_x\sigma_y^2-304\sigma_x^2\sigma_y,$$
$$\mathbf{G}_{N_y}^{-1}\bigotimes\mathbf{G}_{N_x}^{-1}:-32\sigma_x\sigma_y-544\sigma_x\sigma_y^2-688\sigma_x^2\sigma_y,$$
$$\mathbf{G}_{N_y}^{-1}\bigotimes\mathbf{G}_{N_x}^{2}:-8\sigma_x^2\sigma_y,$$
$$\mathbf{G}_{N_y}^{-1}\bigotimes\mathbf{G}_{N_x}^{-2}:-32\sigma_x^2\sigma_y,$$
$$\mathbf{G}_{N_y}^{2}\bigotimes\mathbf{I}_{N_x}:208\sigma_x\sigma_y^2+8\sigma_y^2-144\sigma_y^3,$$
$$\mathbf{G}_{N_y}^{2}\bigotimes\mathbf{G}_{N_x}:16\sigma_x\sigma_y^2,$$
$$\mathbf{G}_{N_y}^{2}\bigotimes\mathbf{G}_{N_x}^{-1}:64\sigma_x\sigma_y^2,$$
$$\mathbf{G}_{N_y}^{-2}\bigotimes\mathbf{I}_{N_x}:-4\sigma_y^2-104\sigma_x\sigma_y^2+72\sigma_y^3,$$

$$\mathbf{G}_{N_y}^{-2}\bigotimes\mathbf{G}_{N_x}:-8\sigma_x\sigma_y^2,$$
$$\mathbf{G}_{N_y}^{-2}\bigotimes\mathbf{G}_{N_x}^{-1}:-32\sigma_x\sigma_y^2.$$

$\mathbf{Q}_3$:

$$\begin{aligned}
\mathbf{I}_{N_y}\bigotimes\mathbf{I}_{N_x}&:8\sigma_x\sigma_y+120\sigma_x^2\sigma_y+120\sigma_x\sigma_y^2,\\
\mathbf{I}_{N_y}\bigotimes\mathbf{G}_{N_x}&:-16\sigma_x\sigma_y-240\sigma_x\sigma_y^2-248\sigma_x^2\sigma_y,\\
\mathbf{I}_{N_y}\bigotimes\mathbf{G}_{N_x}^{-1}&:8\sigma_x\sigma_y+136\sigma_x^2\sigma_y+120\sigma_x\sigma_y^2,\\
\mathbf{I}_{N_y}\bigotimes\mathbf{G}_{N_x}^{2}&:-16\sigma_x^2\sigma_y,\\
\mathbf{I}_{N_y}\bigotimes\mathbf{G}_{N_x}^{-2}&:8\sigma_x^2\sigma_y,\\
\mathbf{G}_{N_y}\bigotimes\mathbf{I}_{N_x}&:-16\sigma_x\sigma_y-240\sigma_x^2\sigma_y-248\sigma_x\sigma_y^2,\\
\mathbf{G}_{N_y}\bigotimes\mathbf{G}_{N_x}&:32\sigma_x\sigma_y+496\sigma_x^2+\sigma_y^2,\\
\mathbf{G}_{N_y}\bigotimes\mathbf{G}_{N_x}^{-1}&:-16\sigma_x\sigma_y-272\sigma_x^2\sigma_y-248\sigma_x\sigma_y^2,\\
\mathbf{G}_{N_y}\bigotimes\mathbf{G}_{N_x}^{2}&:32\sigma_x^2\sigma_y,\\
\mathbf{G}_{N_y}\bigotimes\mathbf{G}_{N_x}^{-2}&:-16\sigma_x^2\sigma_y,\\
\mathbf{G}_{N_y}^{-1}\bigotimes\mathbf{I}_{N_x}&:8\sigma_x\sigma_y+120\sigma_x^2\sigma_y+136\sigma_x\sigma_y^2,\\
\mathbf{G}_{N_y}^{-1}\bigotimes\mathbf{G}_{N_x}&:-16\sigma_x\sigma_y-248\sigma_x^2-272\sigma_x\sigma_y^2,\\
\mathbf{G}_{N_y}^{-1}\bigotimes\mathbf{G}_{N_x}^{-1}&:8\sigma_x\sigma_y+136\sigma_x^2\sigma_y+136\sigma_x\sigma_y^2,\\
\mathbf{G}_{N_y}^{-1}\bigotimes\mathbf{G}_{N_x}^{2}&:-16\sigma_x^2\sigma_y,\\
\mathbf{G}_{N_y}^{-1}\bigotimes\mathbf{G}_{N_x}^{-2}&:8\sigma_x^2\sigma_y,\\
\mathbf{G}_{N_y}^{2}\bigotimes\mathbf{I}_{N_x}&:-16\sigma_x\sigma_y^2,\\
\mathbf{G}_{N_y}^{2}\bigotimes\mathbf{G}_{N_x}&:32\sigma_x\sigma_y^2,\\
\mathbf{G}_{N_y}^{2}\bigotimes\mathbf{G}_{N_x}^{-1}&:-16\sigma_x\sigma_y^2,\\
\mathbf{G}_{N_y}^{-2}\bigotimes\mathbf{I}_{N_x}&:8\sigma_x\sigma_y^2,\\
\mathbf{G}_{N_y}^{-2}\bigotimes\mathbf{G}_{N_x}&:-16\sigma_x\sigma_y^2,\\
\mathbf{G}_{N_y}^{-2}\bigotimes\mathbf{G}_{N_x}^{-1}&:8\sigma_x\sigma_y^2.
\end{aligned}$$

$\mathbf{Q}_4$:

$$\begin{aligned}
\mathbf{I}_{N_y}\bigotimes\mathbf{I}_{N_x}&:1+58\sigma_x+58\sigma_y+924\sigma_x^2,\\
\mathbf{I}_{N_y}\bigotimes\mathbf{G}_{N_x}&:4\sigma_x+80\sigma_x^2+168\sigma_x\sigma_y+2800\sigma_x^2+1704\sigma_x\sigma_y^2-792\sigma_x^3,\\
\mathbf{I}_{N_y}\bigotimes\mathbf{G}_{N_x}^{-1}&:10\sigma_x+272\sigma_x^2+480\sigma_x\sigma_y+7792\sigma_x^2+6384\sigma_x\sigma_y^2-360\sigma_x^3,
\end{aligned}$$

$$
\begin{aligned}
\mathbf{I}_{N_y} \bigotimes \mathbf{G}_{N_x}^{2} &: 4\sigma_x^2 + 104\sigma_x^2\sigma_y - 72\sigma_x^3,\\
\mathbf{I}_{N_y} \bigotimes \mathbf{G}_{N_x}^{-2} &: 16\sigma_x^2 + 416\sigma_x^2\sigma_y - 288\sigma_x^3,\\
\mathbf{G}_{N_y} \bigotimes \mathbf{I}_{N_x} &: 4\sigma_y + 80\sigma_y^2 + 168\sigma_x\sigma_y + 1704\sigma_x^2\sigma_y + 2800\sigma_x\sigma_y^2 - 792\sigma_y^3,\\
\mathbf{G}_{N_y} \bigotimes \mathbf{G}_{N_x} &: 12\sigma_x\sigma_y + 232\sigma_x\sigma_y^2 + 232\sigma_x^2\sigma_y,\\
\mathbf{G}_{N_y} \bigotimes \mathbf{G}_{N_x}^{-1} &: 616\sigma_x^2\sigma_y + 36\sigma_x\sigma_y + 1144\sigma_x\sigma_y^2,\\
\mathbf{G}_{N_y} \bigotimes \mathbf{G}_{N_x}^{2} &: 8\sigma_x^2\sigma_y,\\
\mathbf{G}_{N_y} \bigotimes \mathbf{G}_{N_x}^{-2} &: 32\sigma_x^2\sigma_y,\\
\mathbf{G}_{N_y}^{-1} \bigotimes \mathbf{I}_{N_x} &: 10\sigma_y + 272\sigma_y^2 + 480\sigma_x\sigma_y + 6384\sigma_x^2\sigma_y + 7792\sigma_x\sigma_y^2 - 360\sigma_y^3,\\
\mathbf{G}_{N_y}^{-1} \bigotimes \mathbf{G}_{N_x} &: 36\sigma_x\sigma_y + 1144\sigma_x^2\sigma_y + 616\sigma_x\sigma_y^2,\\
\mathbf{G}_{N_y}^{-1} \bigotimes \mathbf{G}_{N_x}^{-1} &: 132\sigma_x\sigma_y + 2680\sigma_x^2\sigma_y + 2680\sigma_x\sigma_y^2,\\
\mathbf{G}_{N_y}^{-1} \bigotimes \mathbf{G}_{N_x}^{2} &: 128\sigma_x^2\sigma_y,\\
\mathbf{G}_{N_y}^{-1} \bigotimes \mathbf{G}_{N_x}^{-2} &: 32\sigma_x^2\sigma_y,\\
\mathbf{G}_{N_y}^{2} \bigotimes \mathbf{I}_{N_x} &: 4\sigma_y^2 - 72\sigma_y^3 + 104\sigma_x\sigma_y^2,\\
\mathbf{G}_{N_y}^{2} \bigotimes \mathbf{G}_{N_x} &: 8\sigma_x\sigma_y^2,\\
\mathbf{G}_{N_y}^{2} \bigotimes \mathbf{G}_{N_x}^{-1} &: 32\sigma_x\sigma_y^2,\\
\mathbf{G}_{N_y}^{-2} \bigotimes \mathbf{I}_{N_x} &: 16\sigma_y^2 - 288\sigma_y^3 + 416\sigma_x\sigma_y^2,\\
\mathbf{G}_{N_y}^{-2} \bigotimes \mathbf{G}_{N_x} &: 32\sigma_x\sigma_y^2,\\
\mathbf{G}_{N_y}^{-2} \bigotimes \mathbf{G}_{N_x}^{-1} &: 128\sigma_x\sigma_y^2.
\end{aligned}
$$

By definition we get the similar representation of $\mathbf{W}_\varrho$ ($\varrho = 1, 2, 3$) and $\mathbf{W}_4$ is the same as $\frac{1}{4}\mathbf{Q}_4$.
$4\mathbf{W}_1$:

$$
\begin{aligned}
\mathbf{I}_{N_y} \bigotimes \mathbf{I}_{N_x} &: 1 + 54\sigma_x + 54\sigma_y + 1992\sigma_x\sigma_y + 804\sigma_x^2 + 804\sigma_y^2\\
&\quad + 16824\sigma_x^2\sigma_y + 16824\sigma_x\sigma_y^2 + 1080\sigma_x^3 + 1080\sigma_y^3,\\
\mathbf{I}_{N_y} \bigotimes \mathbf{G}_{N_x} &: 12\sigma_x + 328\sigma_x^2 + 504\sigma_x\sigma_y + 7648\sigma_x^2\sigma_y + 6744\sigma_x\sigma_y^2 - 72\sigma_x^3,\\
\mathbf{I}_{N_y} \bigotimes \mathbf{G}_{N_x}^{-1} &: 6\sigma_x + 136\sigma_x^2 + 240\sigma_x\sigma_y + 3424\sigma_x^2\sigma_y + 2784\sigma_x\sigma_y^2 - 504\sigma_x^3,\\
\mathbf{I}_{N_y} \bigotimes \mathbf{G}_{N_x}^{2} &: 20\sigma_x^2 + 440\sigma_x^2\sigma_y - 360\sigma_x^3,\\
\mathbf{I}_{N_y} \bigotimes \mathbf{G}_{N_x}^{-2} &: 8\sigma_x^2 + 176\sigma_x^2\sigma_y - 144\sigma_x^3,\\
\mathbf{G}_{N_y} \bigotimes \mathbf{I}_{N_x} &: 12\sigma_y + 328\sigma_y^2 + 504\sigma_x\sigma_y + 7648\sigma_x\sigma_y^2 + 6744\sigma_x^2\sigma_y - 72\sigma_y^3,
\end{aligned}
$$

$$\mathbf{G}_{N_y}\bigotimes\mathbf{G}_{N_x}:204\sigma_x\sigma_y+3928\sigma_x\sigma_y^2+3928\sigma_x^2\sigma_y,$$
$$\mathbf{G}_{N_y}\bigotimes\mathbf{G}_{N_x}^{-1}:84\sigma_x\sigma_y+1528\sigma_x\sigma_y^2+2008\sigma_x^2\sigma_y,$$
$$\mathbf{G}_{N_y}\bigotimes\mathbf{G}_{N_x}^{2}:160\sigma_x^2\sigma_y,$$
$$\mathbf{G}_{N_y}\bigotimes\mathbf{G}_{N_x}^{-2}:80\sigma_x^2\sigma_y,$$
$$\mathbf{G}_{N_y}^{-1}\bigotimes\mathbf{I}_{N_x}:6\sigma_y+136\sigma_y^2+240\sigma_x\sigma_y+3424\sigma_x\sigma_y^2+2784\sigma_x^2\sigma_y-504\sigma_y^3,$$
$$\mathbf{G}_{N_y}^{-1}\bigotimes\mathbf{G}_{N_x}:84\sigma_x\sigma_y+1528\sigma_x^2\sigma_y+2008\sigma_x\sigma_y^2,$$
$$\mathbf{G}_{N_y}^{-1}\bigotimes\mathbf{G}_{N_x}^{-1}:4\sigma_x\sigma_y+760\sigma_x\sigma_y^2+760\sigma_x^2\sigma_y,$$
$$\mathbf{G}_{N_y}^{-1}\bigotimes\mathbf{G}_{N_x}^{2}:80\sigma_x^2\sigma_y,$$
$$\mathbf{G}_{N_y}^{-1}\bigotimes\mathbf{G}_{N_x}^{-2}:32\sigma_x^2\sigma_y,$$
$$\mathbf{G}_{N_y}^{2}\bigotimes\mathbf{I}_{N_x}:20\sigma_y^2+440\sigma_x\sigma_y^2-360\sigma_y^3,$$
$$\mathbf{G}_{N_y}^{2}\bigotimes\mathbf{G}_{N_x}:200\sigma_x\sigma_y^2,$$
$$\mathbf{G}_{N_y}^{2}\bigotimes\mathbf{G}_{N_x}^{-1}:80\sigma_x\sigma_y^2,$$
$$\mathbf{G}_{N_y}^{-2}\bigotimes\mathbf{I}_{N_x}:176\sigma_x\sigma_y^2-144\sigma_y^3+8\sigma_y^2,$$
$$\mathbf{G}_{N_y}^{-2}\bigotimes\mathbf{G}_{N_x}:80\sigma_x\sigma_y^2,$$
$$\mathbf{G}_{N_y}^{-2}\bigotimes\mathbf{G}_{N_x}^{-1}:32\sigma_x\sigma_y^2.$$

$4\mathbf{W}_2$:

$$\begin{aligned}\mathbf{I}_{N_y}\bigotimes\mathbf{I}_{N_x}:&1+58\sigma_x+54\sigma_y+2168\sigma_x\sigma_y+924\sigma_x^2+804\sigma_y^2\\&+19464\sigma_x^2\sigma_y+19464\sigma_x\sigma_y^2+1512\sigma_x^3+1080\sigma_y^3,\end{aligned}$$
$$\mathbf{I}_{N_y}\bigotimes\mathbf{G}_{N_x}:4\sigma_x+80\sigma_x^2+152\sigma_x\sigma_y+2192\sigma_x^2\sigma_y+1464\sigma_x\sigma_y^2-792\sigma_x^3,$$
$$\mathbf{I}_{N_y}\bigotimes\mathbf{G}_{N_x}^{-1}:10\sigma_x+272\sigma_x^2+416\sigma_x\sigma_y+6416\sigma_x^2\sigma_y+5424\sigma_x\sigma_y^2-360\sigma_x^3,$$
$$\mathbf{I}_{N_y}\bigotimes\mathbf{G}_{N_x}^{2}:4\sigma_x^2+88\sigma_x^2\sigma_y-72\sigma_x^3,$$
$$\mathbf{I}_{N_y}\bigotimes\mathbf{G}_{N_x}^{-2}:16\sigma_x^2+352\sigma_x^2\sigma_y-288\sigma_x^3,$$
$$\mathbf{G}_{N_y}\bigotimes\mathbf{I}_{N_x}:12\sigma_y+328\sigma_y^2+584\sigma_x\sigma_y+9248\sigma_x\sigma_y^2+7944\sigma_x^2\sigma_y-72\sigma_y^3,$$
$$\mathbf{G}_{N_y}\bigotimes\mathbf{G}_{N_x}:44\sigma_x\sigma_y+728\sigma_x\sigma_y^2+1448\sigma_x^2\sigma_y,$$
$$\mathbf{G}_{N_y}\bigotimes\mathbf{G}_{N_x}^{-1}:3368\sigma_x^2\sigma_y+164\sigma_x\sigma_y+3128\sigma_x\sigma_y^2,$$
$$\mathbf{G}_{N_y}\bigotimes\mathbf{G}_{N_x}^{2}:40\sigma_x^2\sigma_y,$$
$$\mathbf{G}_{N_y}\bigotimes\mathbf{G}_{N_x}^{-2}:160\sigma_x^2\sigma_y,$$

$$
\begin{aligned}
\mathbf{G}_{N_y}^{-1}\bigotimes\mathbf{I}_{N_x}:&6\sigma_y+136\sigma_y^2+272\sigma_x\sigma_y+4256\sigma_x\sigma_y^2+3264\sigma_x^2\sigma_y-504\sigma_y^3,\\
\mathbf{G}_{N_y}^{-1}\bigotimes\mathbf{G}_{N_x}:&20\sigma_x\sigma_y+536\sigma_x^2\sigma_y+344\sigma_x\sigma_y^2,\\
\mathbf{G}_{N_y}^{-1}\bigotimes\mathbf{G}_{N_x}^{-1}:&68\sigma_x\sigma_y+1592\sigma_x\sigma_y^2+1304\sigma_x^2\sigma_y,\\
\mathbf{G}_{N_y}^{-1}\bigotimes\mathbf{G}_{N_x}^{2}:&16\sigma_x^2\sigma_y,\\
\mathbf{G}_{N_y}^{-1}\bigotimes\mathbf{G}_{N_x}^{-2}:&64\sigma_x^2\sigma_y,\\
\mathbf{G}_{N_y}^{2}\bigotimes\mathbf{I}_{N_x}:&520\sigma_x\sigma_y^2-360\sigma_y^3+20\sigma_y^2,\\
\mathbf{G}_{N_y}^{2}\bigotimes\mathbf{G}_{N_x}:&40\sigma_x\sigma_y^2,\\
\mathbf{G}_{N_y}^{2}\bigotimes\mathbf{G}_{N_x}^{-1}:&160\sigma_x\sigma_y^2,\\
\mathbf{G}_{N_y}^{-2}\bigotimes\mathbf{I}_{N_x}:&208\sigma_x\sigma_y^2-144\sigma_y^3+8\sigma_y^2,\\
\mathbf{G}_{N_y}^{-2}\bigotimes\mathbf{G}_{N_x}:&16\sigma_x\sigma_y^2,\\
\mathbf{G}_{N_y}^{-2}\bigotimes\mathbf{G}_{N_x}^{-1}:&64\sigma_x\sigma_y^2.
\end{aligned}
$$

$4\mathbf{W}_3$:

$$
\begin{aligned}
\mathbf{I}_{N_y}\bigotimes\mathbf{I}_{N_x}:&1+54\sigma_x+58\sigma_y+2168\sigma_x\sigma_y+924\sigma_y^2+804\sigma_x^2\\
&+19464\sigma_x^2\sigma_y+19464\sigma_x\sigma_y^2+1512\sigma_y^3+1080\sigma_x^3,\\
\mathbf{I}_{N_y}\bigotimes\mathbf{G}_{N_x}:&12\sigma_x+328\sigma_x^2+584\sigma_x\sigma_y+9248\sigma_x^2\sigma_y+7944\sigma_x\sigma_y^2-72\sigma_x^3,\\
\mathbf{I}_{N_y}\bigotimes\mathbf{G}_{N_x}^{-1}:&6\sigma_x+136\sigma_x^2+272\sigma_x\sigma_y+4256\sigma_x^2\sigma_y+3264\sigma_x\sigma_y^2-504\sigma_x^3,\\
\mathbf{I}_{N_y}\bigotimes\mathbf{G}_{N_x}^{2}:&20\sigma_x^2+520\sigma_x^2\sigma_y-360\sigma_x^3,\\
\mathbf{I}_{N_y}\bigotimes\mathbf{G}_{N_x}^{-2}:&8\sigma_x^2+208\sigma_x^2\sigma_y-144\sigma_x^3,\\
\mathbf{G}_{N_y}\bigotimes\mathbf{I}_{N_x}:&4\sigma_y+80\sigma_y^2+152\sigma_x\sigma_y+2192\sigma_x\sigma_y^2+1464\sigma_x^2\sigma_y-792\sigma_y^3,\\
\mathbf{G}_{N_y}\bigotimes\mathbf{G}_{N_x}:&44\sigma_x\sigma_y+728\sigma_y\sigma_x^2+1448\sigma_y^2\sigma_x,\\
\mathbf{G}_{N_y}\bigotimes\mathbf{G}_{N_x}^{-1}:&344\sigma_x^2\sigma_y+20\sigma_x\sigma_y+536\sigma_x\sigma_y^2,\\
\mathbf{G}_{N_y}\bigotimes\mathbf{G}_{N_x}^{2}:&40\sigma_x^2\sigma_y,\\
\mathbf{G}_{N_y}\bigotimes\mathbf{G}_{N_x}^{-2}:&16\sigma_x^2\sigma_y,\\
\mathbf{G}_{N_y}^{-1}\bigotimes\mathbf{I}_{N_x}:&10\sigma_y+272\sigma_y^2+416\sigma_x\sigma_y+6416\sigma_x\sigma_y^2+5424\sigma_x^2\sigma_y-360\sigma_y^3,\\
\mathbf{G}_{N_y}^{-1}\bigotimes\mathbf{G}_{N_x}:&164\sigma_x\sigma_y+3128\sigma_x^2\sigma_y+3368\sigma_x\sigma_y^2,\\
\mathbf{G}_{N_y}^{-1}\bigotimes\mathbf{G}_{N_x}^{-1}:&68\sigma_x\sigma_y+1592\sigma_y\sigma_x^2+1304\sigma_y^2\sigma_x,\\
\mathbf{G}_{N_y}^{-1}\bigotimes\mathbf{G}_{N_x}^{2}:&160\sigma_x^2\sigma_y,
\end{aligned}
$$

$$\begin{aligned}
&\mathbf{G}_{N_y}^{-1}\bigotimes\mathbf{G}_{N_x}^{-2}:64\sigma_x^2\sigma_y,\\
&\mathbf{G}_{N_y}^{2}\bigotimes\mathbf{I}_{N_x}:88\sigma_x\sigma_y^2-72\sigma_y^3+4\sigma_y^2,\\
&\mathbf{G}_{N_y}^{2}\bigotimes\mathbf{G}_{N_x}:40\sigma_x\sigma_y^2,\\
&\mathbf{G}_{N_y}^{2}\bigotimes\mathbf{G}_{N_x}^{-1}:16\sigma_x\sigma_y^2,\\
&\mathbf{G}_{N_y}^{-2}\bigotimes\mathbf{I}_{N_x}:352\sigma_x\sigma_y^2-288\sigma_y^3+16\sigma_y^2,\\
&\mathbf{G}_{N_y}^{-2}\bigotimes\mathbf{G}_{N_x}:160\sigma_x\sigma_y^2,\\
&\mathbf{G}_{N_y}^{-2}\bigotimes\mathbf{G}_{N_x}^{-1}:64\sigma_x\sigma_y^2.
\end{aligned}$$

By analyzing these expressions above, it is clear that when $\sigma>0$, these matrices $\mathbf{W}_\varrho$ ($\varrho=1,2,3,4$) are all positive. Furthermore, the row sums of $\sum_{\varrho=1}^4\mathbf{W}_\varrho$ are the same and are equal to the row sums of $\mathbf{F}$. Hence we get the third property in Theorem 2.9.

## Appendix C. Proofs in Section 3

### C.1. **The proof of Theorem 3.1.**

**Proof:** By definition we have the formula of each block of $\mathbf{F}_1$ as below

$$(\mathbf{F}_1)_{ii}=\begin{pmatrix}\frac{2}{3}+a(\frac{2\tau}{h_i^2}+\frac{8\tau}{h_ih_{i-1}})+c\frac{\tau}{h_i} & \frac{1}{3}-c\frac{\tau}{h_i}\\ \frac{1}{3}+c\frac{\tau}{h_i} & \frac{2}{3}+a\frac{6\tau}{h_i^2}+c\frac{\tau}{h_i}\end{pmatrix},$$

$$(\mathbf{F}_1)_{i,i+1}=\begin{pmatrix}-a\frac{2\tau}{h_i^2} & 0\\ -a\frac{6\tau}{h_i^2}-c\frac{2\tau}{h_i} & 0\end{pmatrix},\ (\mathbf{F}_1)_{i,i-1}=\begin{pmatrix}-a\frac{2\tau}{h_ih_{i-1}} & -a\frac{6\tau}{h_ih_{i-1}}\\ 0 & 0\end{pmatrix}.$$

With the assumption $c\tau/h_i<1/3$, we define the perturbed matrix $\tilde{\mathbf{A}}$ and the corresponding $M$-matrix in Lemma 3.1 as

$$\tilde{\mathbf{A}}=\begin{pmatrix}0 & \frac{1}{3}-c\frac{\tau}{h_i}\\ \frac{1}{3}+c\frac{\tau}{h_i} & 0\end{pmatrix}\ \text{and}\ \mathbf{A}=\mathbf{F}_1-\tilde{\mathbf{A}}$$

It is clear that

$$||\mathbf{A}||_\infty=\frac{1}{3}+c\frac{\tau}{h_i}. \tag{C.1}$$

Furthermore, for any $\tilde{a}_{ij}\neq0$, we always have $a_{i,i+2}a_{i+2,j}\neq0$ where the indexes are all in the sense of mod N. Then we conclude that $B=2$. By arithmetic-geometric inequality, we have

$$\frac{2\tau}{h_i^2}+\frac{8\tau}{h_ih_{i-1}}\geq6\frac{\tau}{h_i^2}\left(4\sqrt{\frac{h_i}{h_{i-1}}}\right).$$

Use the assumption $\frac{h_i}{h_{i-1}}\geq\frac{1}{16}$, then we have

$$b(\mathbf{A})=\frac{2}{3}+6\sigma_{\min}a+c\frac{\tau}{h_{\max}} \tag{C.2}$$

where $\sigma_{\min} = \min_i \tau/h_i^2$ and $h_{\max}$ is the largest mesh size. Note that

$$\begin{aligned}\zeta(\mathbf{A}) = &\max_i \left\{ \left(\frac{2}{3} + a(\frac{2\tau}{h_i^2} + \frac{8\tau}{h_i h_{i-1}}) + c\frac{\tau}{h_i}\right) \Big/ \left(a\frac{2\tau}{h_i^2}\right), \right. \\ &\left. \left(\frac{2}{3} + a(\frac{2\tau}{h_i^2} + \frac{8\tau}{h_i h_{i-1}}) + c\frac{\tau}{h_i}\right) \Big/ \left(a\frac{2\tau}{h_i h_{i-1}}\right) \right\} \\ = &\max_i \left\{ \frac{h_i^2}{3a\tau} + 1 + 4\frac{h_i}{h_{i-1}} + \frac{ch_i}{2a},\ \frac{h_i^2}{3a\tau}\frac{h_{i-1}}{h_i} + \frac{h_{i-1}}{h_i} + 4 + \frac{ch_i}{2a}\frac{h_{i-1}}{h_i} \right\}.\end{aligned} \tag{C.3}$$

Since $c\tau/h_i < \frac{1}{3}$, we have $ch_i < h_i^2/3\tau$. Together with (C.1), (C.2), (C.3) and (3.7), we finish our proof immediately. □

Graduate School, China Academy of Engineering Physics, Beijing 100088, China
*Email address*: yukaichang21@gscaep.ac.cn

Laboratory of Computational Physics, Institute of Applied Physics and Computational Mathematics, Beijing 100088, China and HEDPS, Center for Applied Physics and Technology, and College of Engineering, Peking University, Beijing 100871, China
*Email address*: cheng_juan@iapcm.ac.cn

Institute of Applied Physics and Computational Mathematics, Beijing 100094, China
*Email address*: liu_yuanyuan@iapcm.ac.cn

Division of Applied Mathematics, Brown University, Providence, RI 02912, USA
*Email address*: chi-wang_shu@brown.edu